Caractérisation en modules fonctionnels des protéines ADAMTS-TSL, par approches de phylogénies

Olivier Dennler

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 605
SCIENCE DE LA VIE ET DE LA SANTE
Spécialité : *Génétique, Génomique et Bioinformatique*

Par
Olivier DENNLER

Caractérisation en modules fonctionnels des protéines ADAMTS-TSL, par approches de phylogénies

Thèse présentée et soutenue à Rennes, le 19 Décembre 2022
Unité de recherche :
Institut de Recherche en Santé, Environnement et Travail (IRSET).
Equipe Dymec (Univ Rennes 1, INSERM, EHESP).
Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA).
Equipe Dyliss (Univ Rennes 1, INRIA, CNRS).

Rapporteurs avant soutenance :

Lydie LANE	Co-directrice du groupe CALIPHO, Université de Genève, SIB
Hugues RICHARD	Directeur de recherche Robert Koch Institute, Berlin

Composition du Jury :

Président :	Pierre TUFFERY	Directeur de recherche INSERM, Paris
Examinateurs :	Lydie LANE	Co-directrice du groupe CALIPHO, Université de Genève, SIB
	Hugues RICHARD	Directeur de recherche Robert Koch Institute, Berlin
	Vincent BERRY	Professeur Université Montpellier, LIRMM
	Pierre TUFFERY	Directeur de recherche INSERM, Paris
Dir. de thèse :	Nathalie THÉRET	Directrice de recherche INSERM, Rennes
Co-dir. de thèse :	François COSTE	Chargé de recherche Inria, Rennes

Invité(s) :

Samuel BLANQUART	Chargé de recherche Inria, Rennes
Catherine BELLEANNÉE	Maîtresse de conférence Université de Rennes 1

Table des matières

TABLE DES FIGURES

Liste des tableaux

INTRODUCTION

La motivation de cette thèse est la recherche de nouveaux motifs fonctionnels dans les protéines ADAMTS (***A*** ***D****isintegrin-like* ***A****nd* ***M****etalloproteinase with* ***T****hrombospondin motifs*) et ADAMTSL (ADAMTS-like) qui forment la famille de protéines ADAMTS-TSL, des protéines impliquées dans de nombreuses pathologies (e.g., cancer, fibrose, arthrite et maladies cardio-vasculaires). Comme la majorité des protéines eucaryotes, les ADAMTS-TSL sont des protéines multidomaines dont la caractérisation fonctionnelle reste un défi.

Les approches classiques de prédiction fonctionnelle des ADAMTS-TSL reposent sur l'identification de domaines/motifs conservés sur la base de signatures présentes dans différentes bases de données (e.g., PROSITE, Pfam). Bien que ces approches permettent de prédire des domaines fonctionnels, comme c'est le cas des domaines disintégrine et metalloprotease, et d'en obtenir une caractérisation approximative, elles ne sont pas suffisantes pour expliquer la diversité fonctionnelle de ces différentes protéines. La famille ADAMTS-TSL est une famille de gènes homologues dite multigènes (26 copies paralogues chez l'humain) dont les protéines sont multidomaines et multifonctionnelles. Il est déjà établi que ces différents gènes ont des fonctions différentes dans l'organisme. Nous aimerions trouver des caractéristiques supplémentaires permettant de mieux les dissocier fonctionnellement.

Dans cette thèse, nous avons développé une approche réunissant l'analyse avancée des séquences et des méthodes phylogénétiques pour identifier les régions fonctionnelles dans les protéines. Nous proposons de nous intéresser à un niveau fin de conservations locales de séquences, que nous appelons *modules*, comme des régions ayant subi d'importantes pressions dans le but de maintenir une fonction de la protéine. L'idée ici est de rechercher des spécificités de séquences (les modules) propres à chacun de ces paralogues, et d'en étudier l'histoire évolutive. Nous pourrons alors regrouper les différents modules acquis de manière concomitante chez un même gène ancestral, de manière à former une *signature de modules*. Ces modules peuvent être porteurs de fonctionnalités. Parallèlement, nous recherchons des phénotypes associés à chacune de ces séquences, pour prédire les fonctionnalités des signatures de modules mises en évidence. Nous proposons d'étendre le principe des approches phylogénomiques à la caractérisation des fonctions de ces mo-

dules, de manière à associer une signature de modules et une fonction sur la base de leur présence estimée chez les mêmes ancêtres. Cette approche est basée sur une intégration originale des histoires évolutives des espèces, des gènes, des modules et des phénotypes. Ces différents scénarios permettent la prédiction de *coapparitions modules-phénotypes* chez un ancêtre et révèlent des ensembles de régions de séquences conservées, non contiguës, partagées depuis cet ancêtre. Nous supposons qu'un tel ensemble de modules conservés résulte de pressions de sélection maintenant le phénotype qui a été acquis en même temps.

Nous avons appliqué cette stratégie pour rechercher des modules fonctionnels dans la famille ADAMTS-TSL en utilisant des données d'interaction protéine-protéine comme phénotypes. Sur la base d'un ensemble d'interactions protéine-protéine impliquant les protéines ADAMTS-TSL humaines, nous avons identifié 45 signatures qui apportent un nouvel éclairage sur l'évolution et la spécialisation des interactions ADAMTS-TSL et représentent des cibles thérapeutiques potentielles. Notre méthode est généralisable à d'autres familles de protéines ainsi qu'à d'autres phénotypes, ce qui ouvre de nouvelles perspectives à la caractérisation fonctionnelle des familles de protéines multifonctionnelles et multidomaines.

Le présent manuscrit se décompose en trois parties. La Partie I présente le contexte dans lequel s'inscrit nos travaux de thèse. Nous commencerons par présenter les généralités sur les gènes, protéines et familles de protéines (Chapitre 1). Nous exposerons ensuite les approches de phylogénies moléculaires (Chapitre 2) ainsi que les différentes méthodes de propagation d'annotations fonctionnelles de protéines (Chapitre 3) avant de présenter les protéines ADAMTS-TSL (Chapitre 4). La Partie II regroupe les différentes contributions. Nous y présenterons nos contributions liées à la construction d'un jeu de données de séquences (214 protéines, 9 espèces) puis à l'inférence d'une phylogénie des gènes ADAMTS-TSL qui aboutit à la proposition de l'arbre de références des gènes ADAMTS-TSL (Chapitre 5). Nous exposerons ensuite notre approche d'identification de modules conservés et l'inférence de leurs histoires évolutives (Chapitre 6). Nous présenterons ensuite nos données d'interactions protéine-protéine et l'inférence de leurs histoires évolutives (Chapitre 7). Puis, nous proposerons notre modèle d'évolution conjointe de l'évolution des gènes, des compositions en modules et des interactions des ADAMTS-TSL (Chapitre 8), qui nous permettra de proposer de nouveaux motifs fonctionnels pour les différents gènes ADAMTS-TSL. Nous finirons par discuter de la pertinence de la méthode proposée et des différents résultats obtenus en Partie III.

⇒ **Problématique**
Les descriptions actuelles des protéines ADAMTS-TSL ne suffisent pas à expliquer les spécificités fonctionnelles des 26 protéines humaines

⇒ **Objectif**
Caractériser des régions fonctionnelles chez les protéines ADAMTS-TSL humaines afin d'expliquer leurs similarités et la spécificité fonctionnelle des différents sous-groupes

⇒ **Hypothèse**
Les fonctions et les régions impliquées dans ces dernières évoluent de manière dépendante

⇒ **Approche**
Identifier des régions conservées (modules) et modéliser conjointement l'évolution des espèces, des gènes, des modules (i.e., régions conservées) et des phénotypes

Première partie

Contexte de la thèse

Chapitre 1

Gènes, protéines et famille de protéines

Les constituants universels que sont les nucléotides d'une part, les acides aminés de l'autre, sont l'équivalent logique d'un alphabet dans lequel serait écrite la structure, donc les fonctions associatives spécifiques des protéines.

Jacques Monod
Le Hasard et la Nécessité (1970)

L'ADN est considéré comme « le support de l'information génétique » d'un individu. L'ADN contient des gènes, dont certains sont exprimés sous forme de protéines qui vont assurer la majorité des fonctions d'une cellule, tout en représentant plus de 60 % de son poids sec. Alors que le génome humain est constitué d'environ 20 000 gènes codants pour des protéines, il est estimé que les phénomènes d'épissages alternatifs conduisent à plus de 70 000 protéines différentes. Les modifications post-traductionnelles (e.g., glycosylation, phosphorylation) vont ajouter de la diversité fonctionnelle conduisant à des centaines de milliers de variants protéiques différents [Aeb+18]. La diversité des protéines est aussi issue de différents types de duplications génétiques, résultants de processus évolutifs au cours de milliards d'années d'évolution (i.e., orthologues et paralogues formant des familles). Le but de ce premier chapitre est de présenter les notions de gène, de protéine, de structure, d'interaction, de fonction, d'orthologie, de paralogie et de divergence ainsi que leurs représentations bio-informatique utilisées dans cette thèse dans le but d'étudier la famille de protéines ADAMTS-TSL.

1.1 Des gènes aux protéines

Le génome est l'ensemble du contenu en ADN d'une cellule, il correspond à l'ensemble du contenu génétique d'une espèce ou d'un individu, qu'il soit codant ou non codant, et contient donc l'ensemble des gènes. Nous nous intéressons dans cette section uniquement aux gènes dont l'expression a pour finalité la synthèse de protéines. Nous allons commencer par décrire les processus biologiques impliqués dans l'expression de ce type de gène codant, ses représentations sous forme de séquences, avant de nous intéresser aux bases de données bioinformatiques dans lesquelles ces séquences sont stockées.

1.1.1 Les protéines comme produits de l'expression du gène

Un gène correspond à une séquence de nucléotides (A, C, G, T) qui sera transcrit et à ses différentes séquences régulatrices. La séquence transcrite est composée d'une succession d'introns et d'exons. On appelle *locus* la localisation d'un gène sur le génome. L'expression des gènes est régulée par des éléments régulateurs, tel que le promoteur ou les *enhancers* et *silencers* (pouvant être en amont ou en aval de leur *locus*). La transcription d'un gène en ARN pré-messager, suivie de l'épissage alternatif de ce dernier, produit des *transcrits alternatifs* sous forme d'ARN messagers matures, variant par la collection d'exons retenue dans chacun d'entre d'eux (Figure 1.1). Chacun de ces ARN messagers matures peut être traduit en une protéine. Les différentes protéines ainsi issues d'un même gène sont appelées *protéines isoformes*. L'ensemble des transcrits d'une cellule correspond à son transcriptome, et l'ensemble des protéines traduites correspond à son protéome.

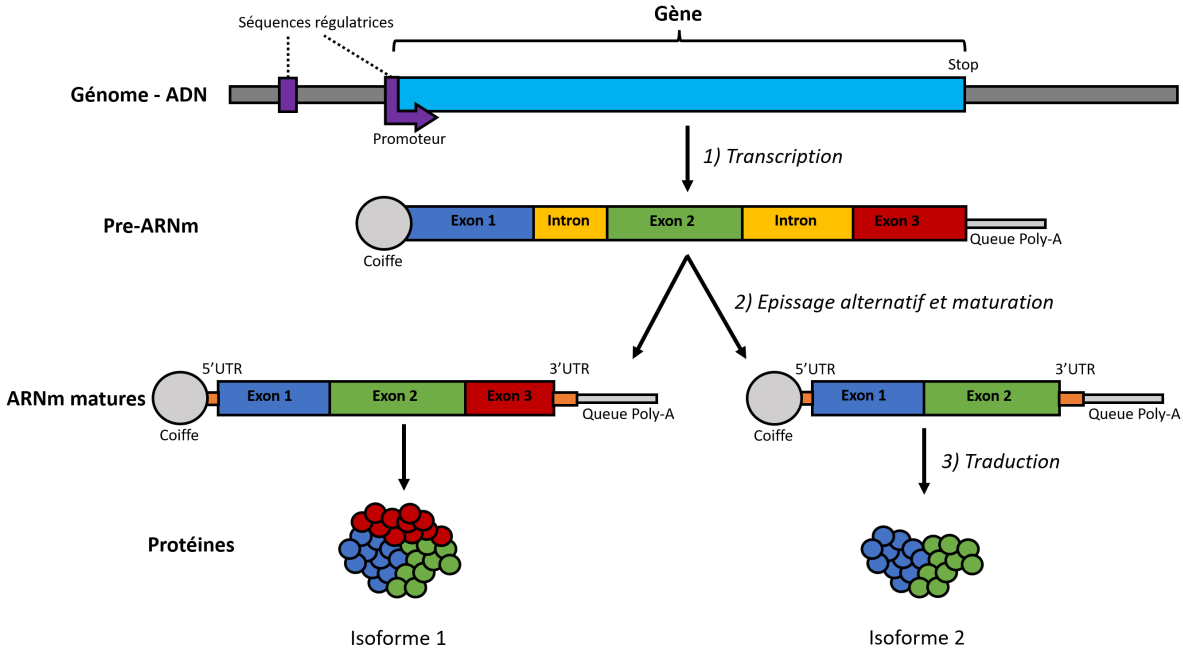

FIGURE 1.1 – **Du gène aux protéines : transcription, épissage alternatif et traduction.** **Un gène** est une séquence de nucléotides située dans le **génome**, son expression est régulée par des séquences régulatrices, comme un promoteur et des éléments localisés en amont ou en aval du *locus* de ce gène (e.g. *enhancer* et *silencer*). Ces éléments régulent la transcription d'un préARNm. Le pré ARNm est constitué d'une suite d'intron et d'exon, l'épissage des introns forme l'ARNm. L'ARNm commence par une région appelée 5'UTR, se poursuit par la région codante (CDS) et se termine par une région 3'UTR. L'expression du gène se déroule en trois grandes étapes. **1) La transcription** correspond à la copie du contenu ADN de ce gène en ARN, son produit étant un ARN pré-messager (**pre-ARNm**). **2) L'épissage alternatif** va exciser de l'ARN pré-messager, les introns ainsi que certains exons, produisant des transcrits alternatifs sous forme d'ARN messagers matures (**ARNm matures**). **3) La traduction** est le nom donné au processus qui permet la synthèse d'une **protéine** à partir de la lecture d'un ARN messager mature.

1.1.2 L'information du gène sous forme de séquences

1.1.2.1 Plusieurs alphabets pour l'information d'un gène

ADN et ARN sont des successions d'acides nucléiques. Il existe 5 nucléotides différents ; Adénine (A), Cytosine (C), Guanine (G), Thymine (T) et Uracile (U). L'Uracile et la Thymine ne sont respectivement pas présents dans l'ADN et l'ARN. L'ADN est une succession de nucléotides ACGT, alors que l'ARN est une succession de nucléotides ACGU. ADN et ARN ont ainsi tous deux un alphabet de taille 4.

Une protéine est une succession d'acides aminés. Les acides aminés incorporés dans une protéine lors de la traduction de l'ARNm sont appelés protéinogènes, on en recense 22, dont 20 standards. Les 2 non standards sont la Pyrrolysine (spécifiques à certains archées) et la Sélénocystéine (spécifique à certaines oxydoréductases). Il existe également des acides aminés non protéinogènes, ils sont pour la plupart absents des protéines et peuvent servir par exemple d'intermédiaires de voies métaboliques. Certains acides aminés non protéinogènes peuvent également être présents dans des protéines, suite à des modifications post-traductionnelles de ces dernières. Dans cette thèse, nous nous intéressons aux protéines comme séquences d'acides aminés protéinogènes standards, ce qui correspond à un alphabet de taille 20.

Un acide aminé est composé d'une partie commune à tous les acides aminés, composée d'un carbone, d'un groupe carboxyle et d'un groupe amine (Figure 1.2). Au sein d'une chaîne peptidique, les acides aminés sont reliés par des liaisons covalentes entre le groupe amine et le groupe carboxyle d'acides aminés successifs. De plus, chaque acide aminé possède une chaîne latérale, différente d'un acide aminé à un autre.

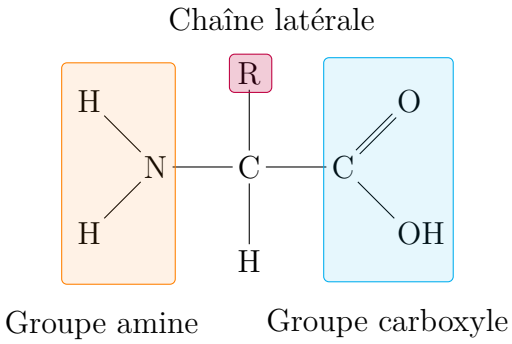

FIGURE 1.2 – **Structure générique d'un acide aminé**

En fonction de la composition de leur chaine latérale (Figure 1.3), les acides aminés vont présenter des propriétés physico-chimiques différentes. En effet, certains vont être composés d'une chaîne latérale polaire (i.e., hydrophile), d'autre d'une chaîne latérale apolaire (i.e., hydrophobe). Les acides aminés basiques vont également posséder des groupes amines (NH2) dans leur chaine latérale, ayant tendance à se charger positivement (en fonction du pH). À l'inverse, les acides aminés acides vont posséder un groupe carboxyle (COOH), pouvant se charger négativement (en fonction du pH). Cette variété de caractéristiques physico-chimiques va engendrer des interactions chimiques diverses entre les différents acides aminés et leur milieu (e.g., repliement de la protéine, solubilité, interactions, réactions biochimiques). Par l'effet de la sélection naturelle, les acides aminés de propriétés communes sont en général plus fréquemment substitués les uns par les autres durant l'évolution de la protéine.

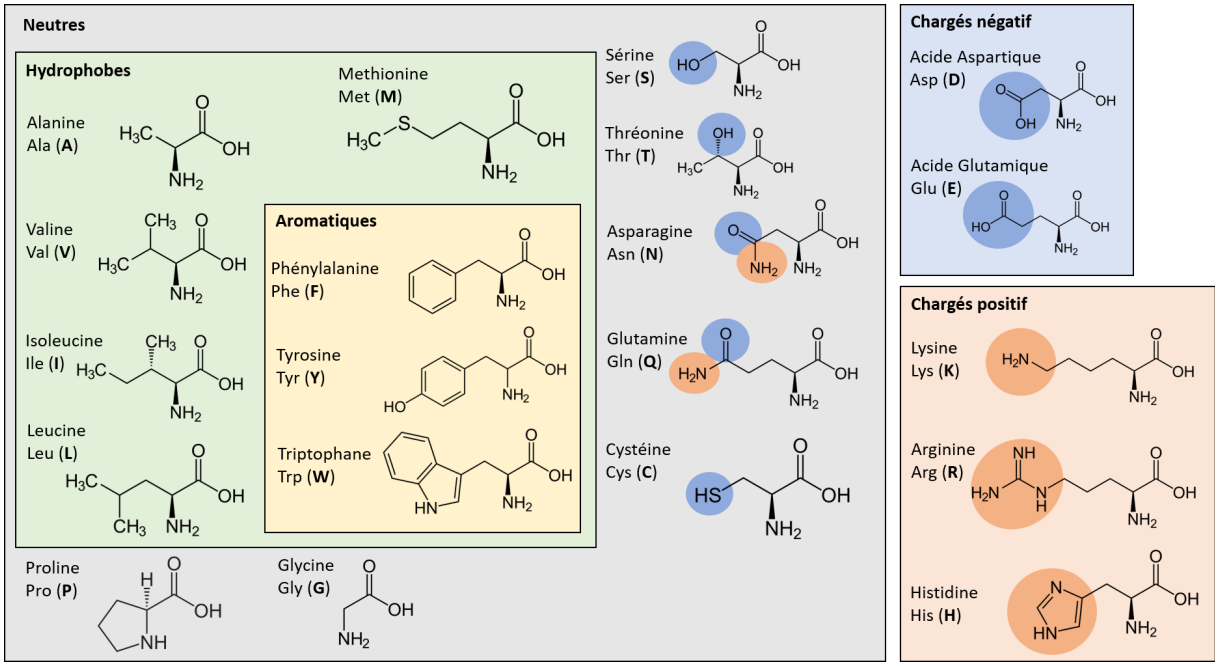

FIGURE 1.3 – **Structure des 20 L-acides aminés protéinogènes standards**

1.1.2.2 Traduction et code génétique : passage d'un alphabet à un autre

Le passage d'une séquence de nucléotides (gène, transcrit) à une séquence d'acide aminé (protéine) se fait au cours de la traduction, qui biosynthétise la chaine d'acide aminé à partir de la séquence en acide aminé d'un ARN messager mature. Pour ceci, les acides nucléiques sont lus par codons (succession de trois nucléotides ou triplet de nucléotides), de manière à synthétiser une protéine, acide aminé par acide aminé. L'association d'un codon à un acide aminé est ce qu'on appelle le code génétique (Figure 1.4). Les 64 codons possibles sont tous associés à un acide aminé parmi les 20 protéinogènes et à trois codons « stop » permettant de terminer la traduction. Cela signifie qu'un acide aminé peut être codé par plusieurs codons différents.

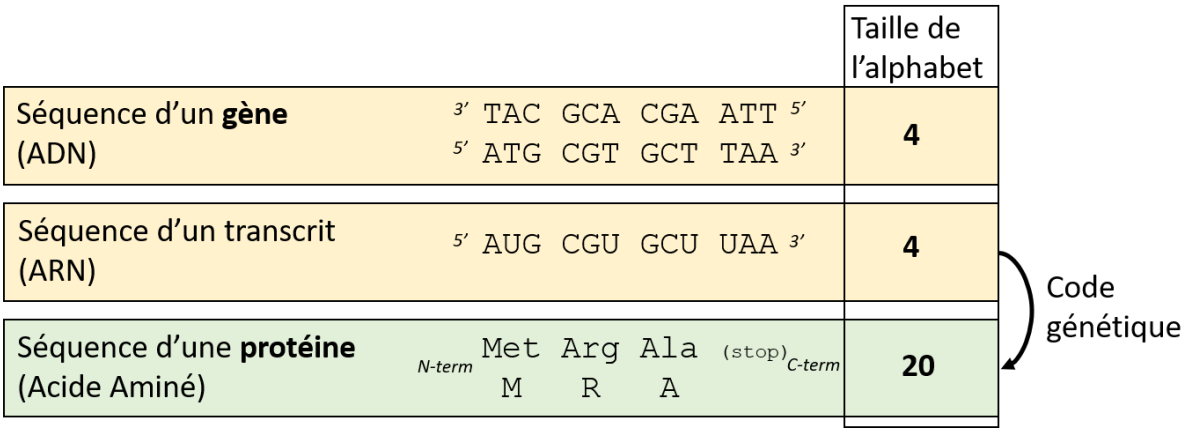

FIGURE 1.4 – **Séquences génétiques**
Le code génétique permet de traduire une séquence de nucléotides (alphabet de taille 4) en une séquence de protéines (alphabet de taille 20).

1.1.3 Les bases de données de séquences biologiques

Il existe un grand nombre de bases de données de séquences biologiques [CHW17]. Le but de cette section est de présenter les bases de données de séquences génétiques utilisées dans cette thèse.

1.1.3.1 DDBJ, ENA et GenBank : une collaboration entre bases de données de nucléotides

La DDBJ (*DNA Data Bank of Japan*) [Kod+15], l'ENA (*European Nucleotide sequence database*) [Kul+07] et GenBank (*GenBank nucleotide sequence database*) [Aga+18] sont trois bases de données de nucléotides, gérées par des organismes différents, faisant partie de l'INSDC (*International Nucleotide Database Collaboration*). Cette collaboration permet l'échange de données de manière très régulière entre les soumissions faites dans chacune de ces bases de données afin que les bases de données DDBJ, ENA et GenBank présentent strictement le même contenu. En mars 2022, elles contenaient 2 688 127 662 séquences, pour un total de 17 470 756 718 739 bases.

1.1.3.2 CCDS : un consensus du codant

La base de données CCDS (*Consensus Coding Sequence*) [Far+14] regroupe les informations convergentes sur les régions codantes (CDS) de l'humain et la souris. Différentes ressources (e.g., transcrits alternatifs, protéines isoformes, annotation de génome) peuvent présenter des informations différentes à propos d'un même gène et sa séquence codante, le but du projet CCDS est de regrouper les informations convergentes au sein d'un consensus. Les protéines, transcrits et gènes annotés différemment et issus de différentes ressources sont regroupés sur la base de leur région codante. C'est ainsi que CCDS met à disposition des consensus de bonne qualité des leurs positions génomiques ainsi que des transcrits alternatifs associés, et cela, sur la base de *curation* manuelle par des experts. En Juin 2022, CCDS contenait 19 030 gènes humains et 20 486 gènes murins.

1.1.3.3 RefSeq : une base de données controlée

La base de données RefSeq [PTM07] du NCBI (*National Center for Biotechnology Information*) recense toutes les séquences avec un identifiant unique des génomes, des transcrits et des protéines pour tous types d'organismes (archées, bactéries, eucaryotes, virus). Les séquences issues de GenBank sont contrôlées de manière à avoir une entrée

par séquence et sont ensuite annotées avec des informations telles que les domaines, les variations génétiques et les entrées bibliographiques correspondantes. La totalité des informations présentes dans RefSeq sont facilement accessibles, que ce soit par le site, les recherches via l'outil d'alignement `BLAST`, le serveur ftp ou par accès de programmation. En mai 2022 (*release 212*) la base de données RefSeq contenait 229 417 182 protéines, 44 805 833 transcrits, correspondant à 119 373 organismes.

1.1.3.4 UniprotKB, Swiss-prot

UniprotKB et Swiss-prot sont deux bases de données de séquences protéiques du consortium Uniprot, constitué d'équipes de l'EBI (*European Bioinformatics Institute*), du SIB *Swiss Institute of Bioinformatics*) et du PIR (*Protein Information Ressource*). UniprotKB est considérée comme la ressource centrale pour les séquences protéiques et leurs annotations fonctionnelles. UniprotKB est divisée en deux grandes sections : 1) UniProtKB/Swiss-Prot [BA99] et 2) UniProtKB/TrEMBL [Bat+15]. La différence entre ces deux sections se situe au niveau de la qualité des annotations : les entrées Swiss-Prot sont validées par curation manuelle et annotées à base d'informations extraites de la littérature, alors que les entrées TrEMBL sont annotées automatiquement par des méthodes computationnelles. Début 2022, TrEMBL contenait 231 354 261 séquences alors que Swiss-Prot en contenait 567 483. La majorité des entrées de UniprotKB sont annotées avec leurs domaines protéiques, leurs isoformes, leur localisation cellulaire, leur expression tissulaire, leurs structures connues ainsi que d'autres informations connues associées aux différentes entrées.

1.1.3.5 Liens et spécificités des bases de données

Chaque base de données contient différentes informations, que ce soit par leur type, leur quantité et les problématiques dans lesquelles elles s'ancrent. En effet, une base de données ne se résume pas à un stockage de données, mais vise à répondre à un besoin. Que ce soit le stockage d'un grand nombre de données brutes (e.g., DDBJ, ENA, GenBank), le besoin de connecter des données de manière à gagner des plus-values (e.g., CCDS) ou de les annoter en les connectant à différents types de connaissances (e.g., Swiss-Prot). La diversité d'information que présente chaque base de données pour potentiellement parler d'un même gène est tant une richesse qu'un problème à surmonter. Permettre de relier les informations de chaque base de données, sans pour autant perdre les spécificités de chacune, est faisable via les références croisées (*Cross-References*). Les entrées d'une base

de données sont associées à un identifiant qui lui est propre, mais elles présentent aussi les identifiants d'autres bases de données qui décrivent la même entité biologique.

1.2 Séquence, structure et fonction d'une protéine

1.2.1 Repliements et structure d'une protéine

Les caractéristiques physico-chimiques des acides aminés et du milieu conditionnent le repliement de la séquence de la protéine en une structure tridimensionnelle. C'est sous cette forme que la protéine interagit avec les autres molécules et exerce ainsi ses fonctions.

1.2.1.1 Quatre niveaux de repliement

La structure d'une protéine est généralement décrite via quatre niveaux de repliement (Figure 1.5), qui expliquent le passage d'une séquence de résidus à une structure tridimensionnelle.

- **Structure primaire**
  La séquence d'une protéine correspond à son niveau primaire de structure, c'est-à-dire à la liaison des acides aminés successifs par une liaison covalente : la liaison peptidique.
- **Structure secondaire**
  Les acides aminés d'une protéine peuvent également interagir de façon non covalente, par liaison hydrogène, provoquant le repliement local de régions de la protéine. Ces repliements constituent la structure secondaire d'une protéine, les hélices alpha et les feuillets beta étant les structures secondaires les plus communes.
- **Structure tertiaire**
  Les propriétés physico-chimiques des différents acides aminés influencent le repliement de la protéine par des interactions hydrophobes, des liaisons ioniques, des liaisons de Van der Waals ainsi que des liaisons covalentes (e.g., pont disulfure). Le repliement stable de la protéine, qui en résulte, correspond à ce qu'on appelle la structure tertiaire de la protéine.
- **Structure quaternaire**
  Il existe des protéines dites multimériques, elles sont composées de plusieurs chaînes protéiques, chacune de ces chaînes correspond à une sous unité de la protéine multimérique. Chaque sous unité possède sa propre structure tertiaire, et l'agencement de toutes les sous unités d'une protéine multimérique met en œuvre des liaisons non covalentes et correspond à la structure quaternaire de la protéine.

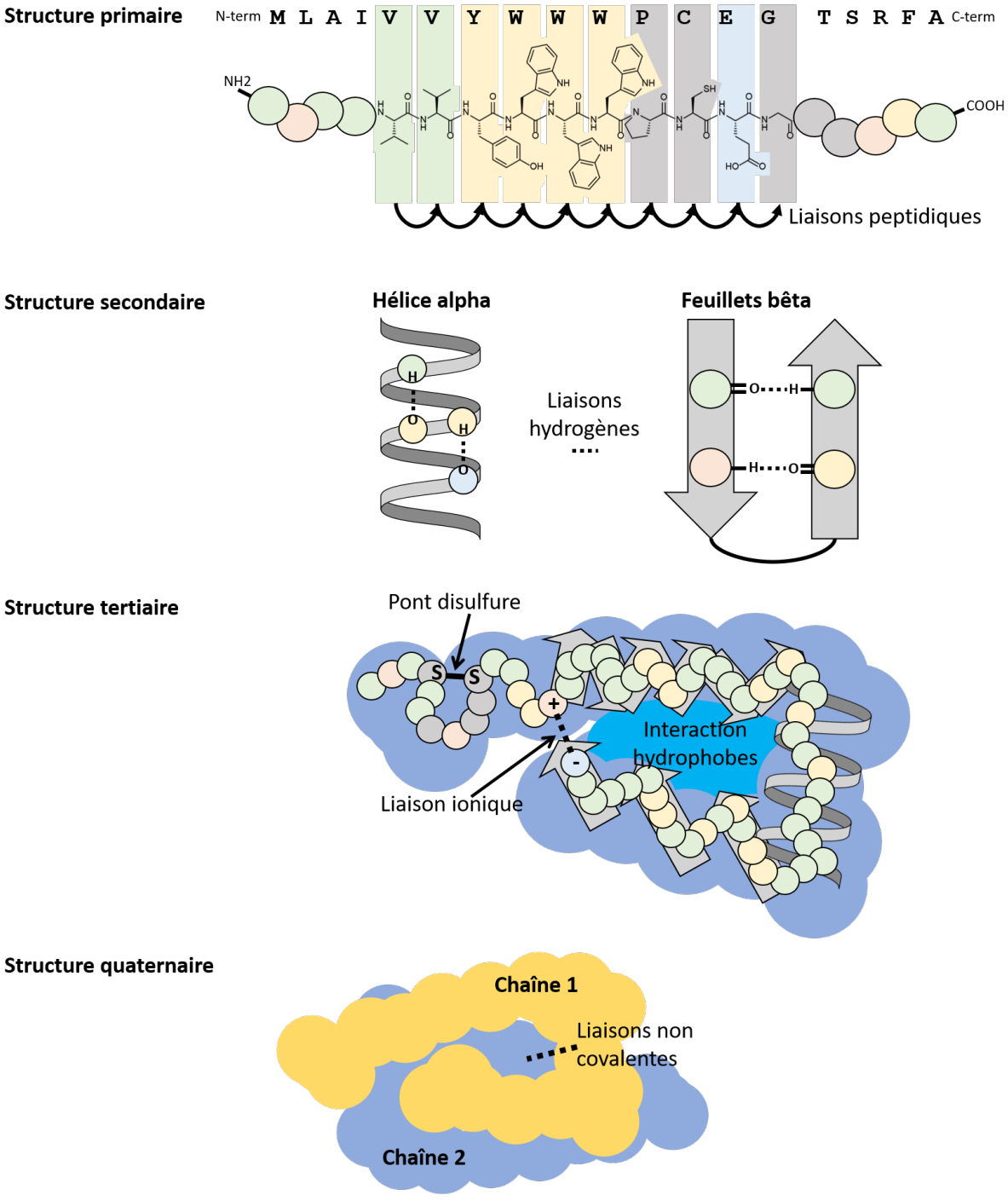

FIGURE 1.5 – **Les quatre niveaux de repliement d'une protéine**

1.2.1.2 Obtention d'une structure de protéine

Il existe différentes méthodes expérimentales permettant de capturer des informations sur la structure d'une protéine. Les plus répandues sont la cristallographie à rayons X, la spectroscopie RMN (Résonance Magnétique Nucléaire) ou la cryo-microscopie électronique (et ses dérivées). Ces méthodes permettent d'obtenir des informations sur la position tridimensionnelle de chaque atome de la protéine à un instant précis. Les coordonnées cartésiennes de chaque atome composant la protéine sont écrites dans un fichier au format .pdb, qui décrit ainsi la structure tridimensionnelle d'une protéine. La *Protein Data Bank* (PDB) est la base de donnée qui regroupe la majorité de ces structures [Bur+21]. Cependant, purifier, isoler, cristalliser une protéine afin d'en obtenir la structure nécessite une démarche expérimentale longue et coûteuse et peut se révéler très complexe pour certaines protéines. Par conséquent, de nombreuses protéines ne possèdent pas de structure PDB ou n'ont pas de structures complètes [DC15] (Figure 1.6).

Le coût, le temps et la complexité que représentent les approches expérimentales d'une part, et l'intérêt colossal que de telles structures représentent pour la pharmacologie, a motivé le développement d'approches prédictives de structure de protéine. Et bien que le problème soit complexe, l'information nécessaire est présente dans la séquence primaire de la protéine. L'idée générale est de faire le lien entre l'information de la séquence et la structure qui en résulte, en se basant sur les structures déjà résolues stockées dans la PDB.

C'est ainsi que depuis presque trois décennies, des scientifiques développent des outils et programmes ayant pour but de prédire la structure d'une protéine, sur la seule base de sa séquence. `I-TASSER` [Yan+15] est un des premiers programmes permettant de prédire la structure d'une séquence *ab initio*, par recherche de séquences similaires dont des structures sont disponibles dans la PDB. Mais l'avancée récente la plus notable dans le domaine de la prédiction de structures revient à `AlphaFold` [Jum+21]. Développée par Deepmind depuis 2018, `AlphaFold` est un logiciel d'intelligence artificielle de prédiction de structure 3D de protéines, nécessitant uniquement des alignements multiples contenant la séquence de la protéine (Figure 1.6). Les prédictions d'`AlphaFold` ont permis la mise à disposition des structures complètes du protéome de l'humain, ainsi que de 47 autres organismes [Var+22].

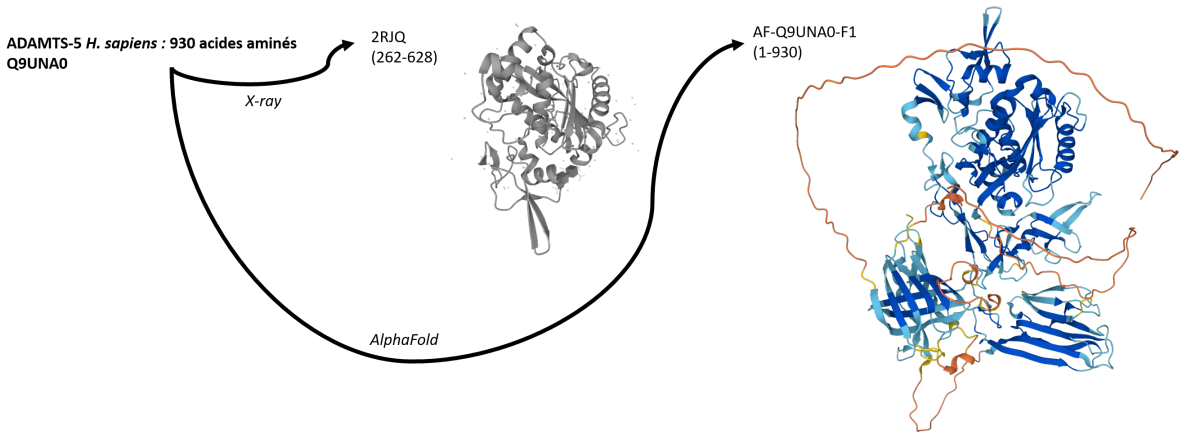

FIGURE 1.6 – **Structures obtenues par rayons X et AlphaFold de la protéine ADAMTS-5** Structure 2RJQ d'ADAMTS-5, obtenue expérimentalement par rayons X. La protéine ADAMTS-5 possède 930 résidus, or seuls 366 acides aminés (de 262 à 628) ont pu être cristallisés dans cette structure. La structure AF-Q9UNA0-F1 prédite par AlphaFold à partir de sa séquence protéique Q9UNA0 propose des coordonnées pour les 930 résidus. Les résidus sont colorés en fonction de la confiance du modèle envers leurs positions.

1.2.2 Les briques de base de l'évolution des protéines

Nous pouvons faire une analogie dans laquelle la séquence d'une protéine peut être vue comme une phrase dont le sens sera la ou les fonctions de la protéine. Les domaines étant les mots composant cette phrase [Yu+19]. Ces domaines étant construits sur des segments plus conservés, qui comme des syllabes seraient partagées par divers mots/domaines.

1.2.2.1 Domaine et organisation modulaire des protéines

Au sein des structures des protéines, les scientifiques ont identifié des briques de base structurale, appelées domaines. Un domaine est une unité structurale minimale autonome de repliement de la séquence d'acides aminés. L'organisation modulaire des protéines [Doh+20] correspond à la réutilisation d'un nombre limité de domaines au sein d'une très grande variété de protéines [KWK02]. En effet, il existe un nombre limité de repliements possibles pour une séquence en acide aminés et la diversité des protéines repose sur la réutilisation de domaines comme briques de bases d'évolution [Doo95]. De plus, 67 % des protéines Eucaryotes sont des protéines multidomaines [Mar+06]. L'accumulation de ces différents domaines, leur répétition et la combinatoire engendrée contribuent globalement à la fonction de la protéine (comme illustré en Figure 1.7 pour le domaine PH qui se retrouvent dans différentes protéines, aux combinatoires en domaines différentes).

1.2.2.2 Segment de sous-domaines comme point de départ des domaines

Bien que les domaines soient généralement considérés comme les briques ou les segments de bases des protéines, de récentes études ont mis en lumière la présence de segments récurrents plus petits que les domaines [Kol+21 ; Kol21]. Ces segments sont des sous unités de domaines présents dans un grand nombre de domaines, contextes et architectures différents. Les domaines seraient ainsi constitués sur la base de segments plus petits, conservés et partagés par d'autres domaines. Le reste du domaine étant variable et différent d'un domaine à l'autre. À la différence des domaines qui sont définis par un repliement structural, ces segments de sous-domaines ne sont pas forcément assez longs pour avoir un repliement autonome.

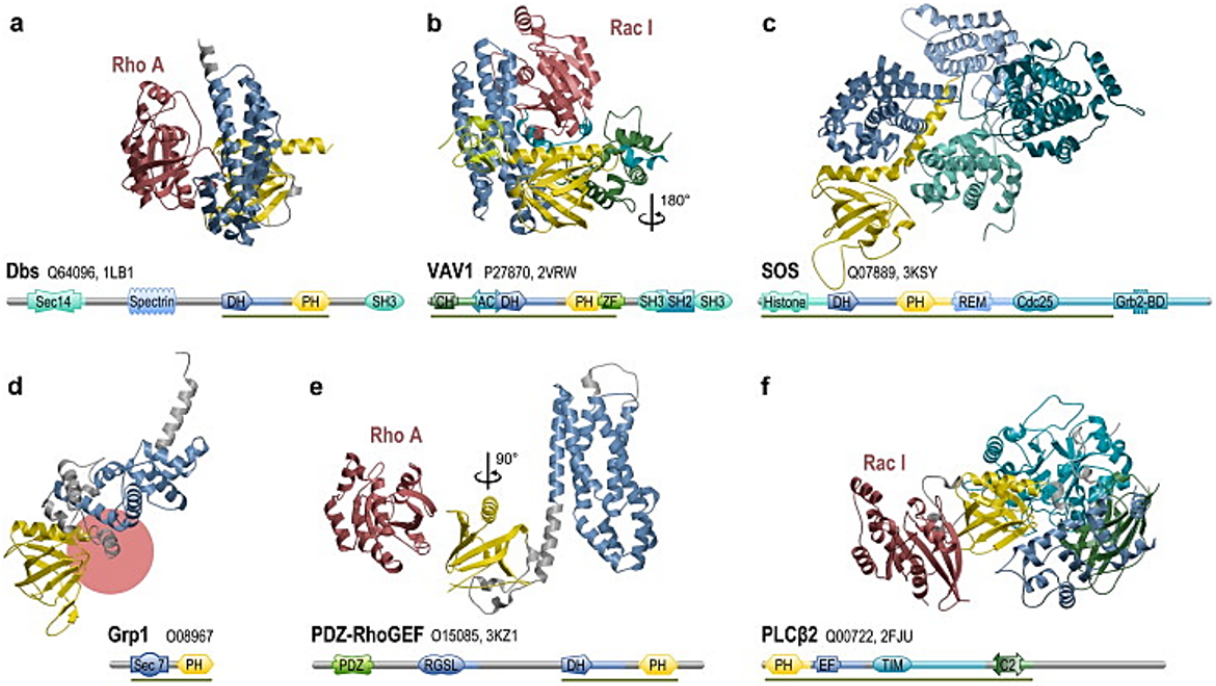

FIGURE 1.7 – **Exemple de réutilisation de domaine au sein de protéines différentes**, d'après [SW12].
Le domaine PH (*Pleckstrin homology*) est représenté en jaune au sein de six protéines différentes. Ces six protéines possèdent une combinatoire en domaine différentes.

1.2.3 Interaction protéine-protéine

1.2.3.1 La structure d'une protéine comme discriminant de ses interactions

Une protéine peut interagir et se fixer spécifiquement avec différentes molécules et macromolécules (Figure 1.8). Le partenaire d'interaction d'une protéine s'appelle un ligand, une protéine en ayant généralement plusieurs possibles. La capacité d'interaction d'une protéine et sa spécificité pour différents ligands définit les fonctions de la protéine. L'interaction d'une protéine avec son ligand se fait par le biais d'une surface de contact, composée par différents résidus qui lui permettent de se lier chimiquement avec son partenaire. Dans un grand nombre d'interactions, un ou plusieurs domaines permettent l'interaction avec le partenaire. Dans le cadre des interactions protéine-protéine, l'interaction peut également être décrite comme domaine-domaine [Den+02], où un domaine d'une protéine interagit spécifiquement avec un domaine de son partenaire. En raison des repliements de la chaîne protéique, les résidus impliqués dans l'interaction peuvent être éloignés dans la séquence de la protéine, tout en étant proche dans sa structure.

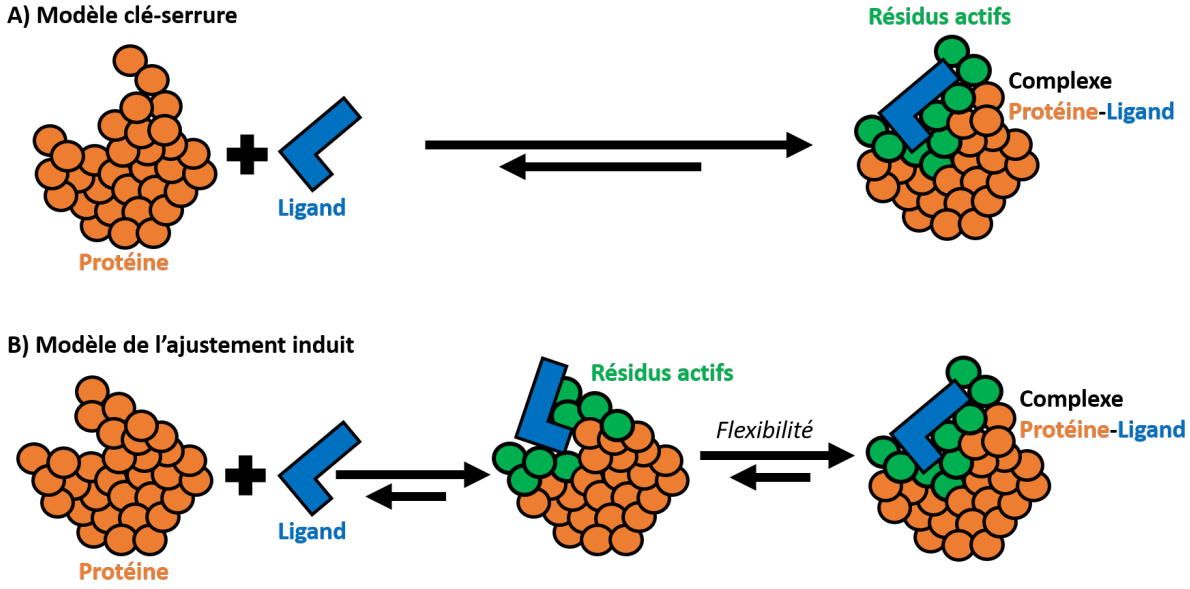

FIGURE 1.8 – **Interaction Protéine-Ligand.**
Représentation schématique des deux modèles d'interaction Protéine-Ligand. La protéine est représentée en orange, ses résidus impliqués dans l'interaction en vert, et son ligand en bleu. **A) Modèle clé-serrure**, les interfaces d'interaction de la protéine et de son ligand sont complémentaires, ce qui leur permet de constituer un complexe. **B) Modèle de l'ajustement induit**, les structures de la protéine et/ou de son ligand se déforment afin de permettre à leurs interfaces d'interaction de s'ajuster.

1.2.3.2 Enzymes et pseudoenzymes

Les enzymes sont des protéines possédant une propriété catalytique. C'est-à-dire qu'elles ont la capacité de modifier la cinétique d'une réaction chimique (e.g., en abaisser l'énergie d'activation). Une réaction catalysée par une enzyme prend en entrées des molécules que l'on appelle substrats, et crée un ou plusieurs produits, l'enzyme n'étant pas modifiée suite à la réaction. La grande particularité des enzymes comparées aux autres catalyseurs est leur très grande spécificité vis-à-vis de leurs substrats, et donc des réactions qu'elles catalysent. En effet, comme toute protéine, les enzymes possèdent une structure tridimensionnelle qui leur est propre, leur permettant d'interagir spécifiquement avec des substrats. Le modèle accepté actuellement pour la liaison enzyme-substrat est le modèle de l'ajustement induit (Figure 1.8). Ce modèle décrit la structure de l'enzyme comme flexible, remodelant en permanence son site actif, de manière à créer une liaison substrat-enzyme stable. Il s'oppose à l'ancien modèle clé-serrure qui considérait la structure de l'enzyme et de ses substrats comme stable et ainsi complémentaire, ce qui permettait leur liaison.

Cependant, il existe aussi ce qu'on appelle des pseudoenzymes [TOT02 ; MFE17 ; Rib+19]. Une pseudoenzyme est une protéine très similaire d'un point de vue séquence/structure à une enzyme (i.e., généralement une copie paralogue), mais ne possédant pas d'activité catalytique. Elles réalisent alors des fonctions cellulaires importantes, comme la régulation de l'activité des enzymes auxquelles elles ressemblent [EM16].

1.2.3.3 PSICQUIC : un accès standardisé et mutualisé aux bases de données d'interactions moléculaires

Les Interactions Protéine-Protéine (PPI pour *Protein-Protein Interaction*) peuvent être détectées par différentes méthodes expérimentales. Les plus classiques sont la technique de double hybride et l'analyse par spectrométrie de masse de complexes purifiés qui permettent de mettre en évidence une liaison physique entre protéines. Il existe d'autres méthodes et d'autres protocoles biochimiques permettant d'étudier des PPI, certaines permettant même d'identifier les régions ou les résidus des séquences qui sont suffisants et nécessaires à l'interaction, par exemple en utilisant des protéines chimériques [Adr+19] (si une région est absente de la chimère et que l'interaction est perdue, cette région est importante pour l'interaction).

Comme pour chaque type de données biologiques, il existe un grand nombre de bases de données d'interactions, plus ou moins spécifiques vis-à-vis des méthodes de détection ou des organismes qu'elles regroupent. Le but ici n'est pas d'en faire un catalogue, mais plutôt de présenter un service web qui permet d'interroger la majorité d'entre elles en une fois : PSICQUIC [Ara+11]. Le service PSICQUIC (*PSI Common Query Interface*) est un effort du consortium HUPO-PSI (*HUPO Proteomics Standard Initiative*) afin de standardiser l'accès aux différentes bases de données de PPI (Figure 1.9). En 2022, PSICQUIC permet l'accès à 34 bases de données d'interactions différentes, sans les centraliser (chaque base de données reste indépendante), mais avec comme contrainte que les PPI y soient décrites au format MITAB (*Molecular Interaction Table*) [Per+19], ce qui permet d'interroger indépendamment les 34 bases de données avec une seule requête, dont les réponses sont regroupées. Une requête PSICQUIC se fait à partir d'un identifiant (e.g., UniprotKB) et son résultat est un fichier au format MITAB standardisé prenant en compte les particularités des différentes bases de données dont les résultats sont issus. Différentes bases de données peuvent partager une même interaction, qu'elle soit issue de la même publication, ou issue de publications différentes. À noter que de nombreuses bases de données utilisent différents identifiants pour nommer les différents partenaires (e.g., identifiants uniprots, identifiants refSeq). De manière à réduire la redondance des bases de données et synchroniser les identifiants, PSICQUIC possède une fonction de regroupement des réponses d'une requête.

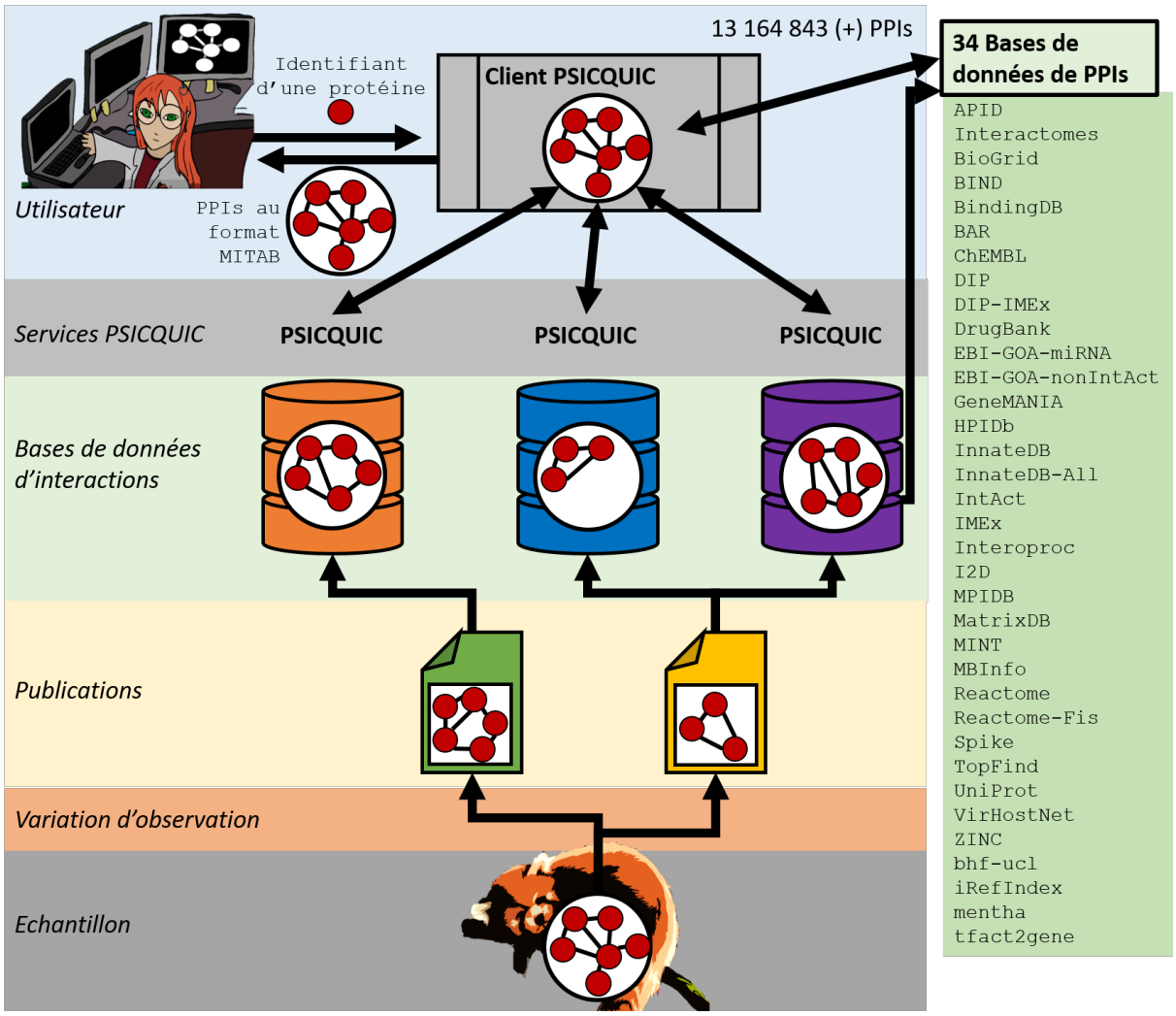

FIGURE 1.9 – **L'interface PSICQUIC (PSI Common Query Interface).**
Représentation de l'interface PSICQUIC, un utilisateur peut en interroger le client avec l'identifiant(s) de protéine(s) d'intérêt(s), qui va alors interroger les 34 bases de données d'interactions pour récupérer les PPIs impliquant la protéine d'intérêt et les retourner au format MITAB. Adapté depuis la documentation du `Github` de PSICQUIC.

1.3 Famille de protéines homologues

Nous allons maintenant nous intéresser aux familles de protéines homologues, à la manière dont elles sont issues de duplications depuis un ancêtre commun, et à la manière dont les copies ont divergé sur le plan séquence/structure, avant de nous intéresser aux manières permettant de détecter et de regrouper des protéines issues d'une origine évolutive commune.

1.3.1 Évolution de protéines homologues

1.3.1.1 Protéines homologues et ancêtre commun

Une famille de protéines homologues correspond à un ensemble de protéines possédant une origine évolutive commune. On parle également de famille de gènes, les protéines étant leurs produits. En effet, un gène ancestral, présent dans un génome ancestral, peut subir différents types de duplication au cours de l'évolution (Figure 1.10). Un gène ancestral peut être dupliqué au cours des divergences du génome dont il provient. Un évènement de spéciation d'un génome ancestral a pour conséquence deux génomes différents, dont chacun possède une copie du gène ancestral. Ces deux copies sont appelées copies orthologues (deux génomes, une copie par génome). Il est également possible qu'un gène ancestral soit dupliqué au sein d'un génome donné, ce qui donne deux copies paralogues (deux copies au sein d'un même génome).

Des fragments génomiques peuvent être dupliqués au cours de différents évènements, que ce soit par duplication de régions chromosomiques (e.g., duplication d'un fragment suite à une réparation non homologue de l'ADN, translocation chromosomique), ou par duplication de régions génomiques de plus petites tailles (e.g., crossing-over inégal, retrotransposition d'éléments transposables, échange ectopique, transfert horizontal). Il existe aussi des évènements de duplications complètes du génome au sein d'une lignée.

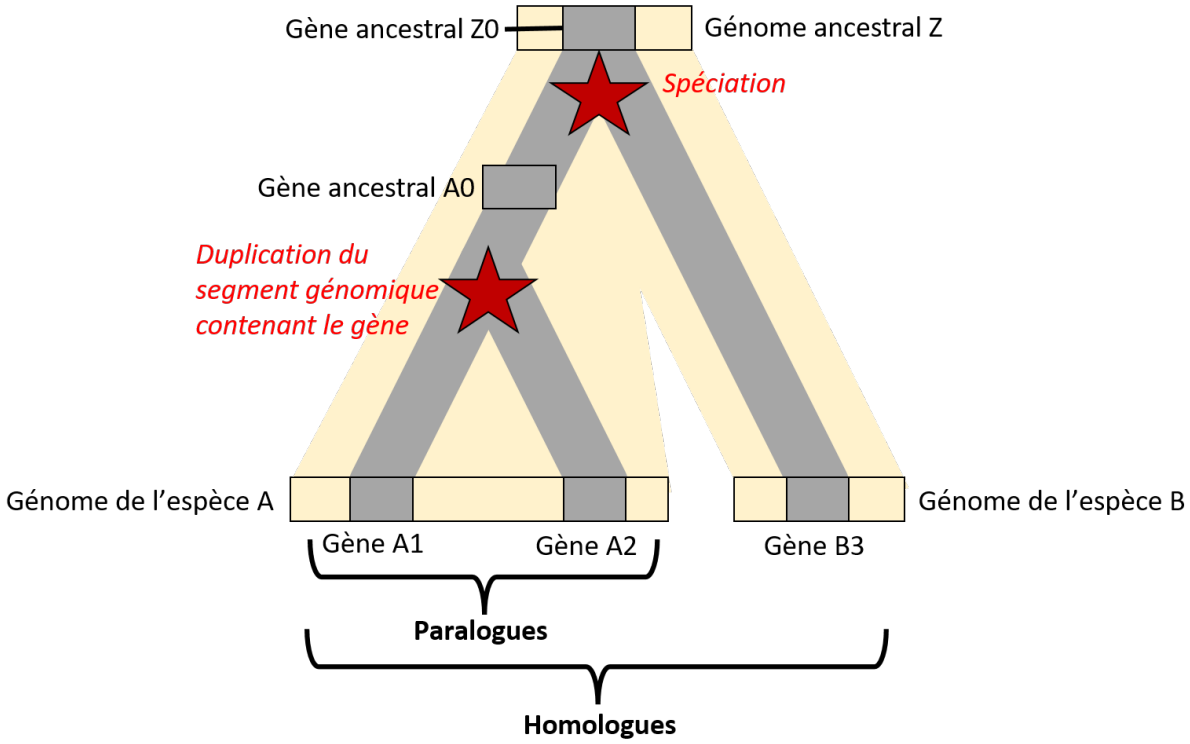

FIGURE 1.10 – **Famille de protéines homologues**
Un génome ancestral Z contenant un gène ancestral Z0 subit une spéciation qui a pour conséquence deux génomes différents : celui de l'espèce A et celui de l'espèce B. Les deux génomes issus de cette spéciation possèdent chacun une copie orthologue du gène : A0 et B3. Au sein de l'espèce A, une duplication d'un segment génomique a pour effet la duplication du gène en deux copies paralogues : A1 et A2. Les gènes A1, A2 et B3 sont alors des gènes homologues issus d'un même gène ancestral (Z0), qui expriment des protéines dites homologues.

1.3.1.2 Copie et divergence de séquences et de fonctions

Au cours de l'évolution, les séquences génétiques varient. Le génome, et donc les gènes, peuvent acquérir des mutations ponctuelles (i.e., substitutions). Il y a alors divergence de leurs séquences (Figure 1.11). C'est pourquoi, bien qu'issus d'un même gène ancestral, deux gènes d'une même famille (et leur produit protéique) différent d'un point de vue séquence. Au cours du temps, deux copies issues d'un même gène ancestral, par accumulation de mutations totalement aléatoires (i.e., substitutions ponctuelles), ou par duplication/transfert/perte de segments génomiques en leur sein, vont ainsi voir leur séquence diverger.

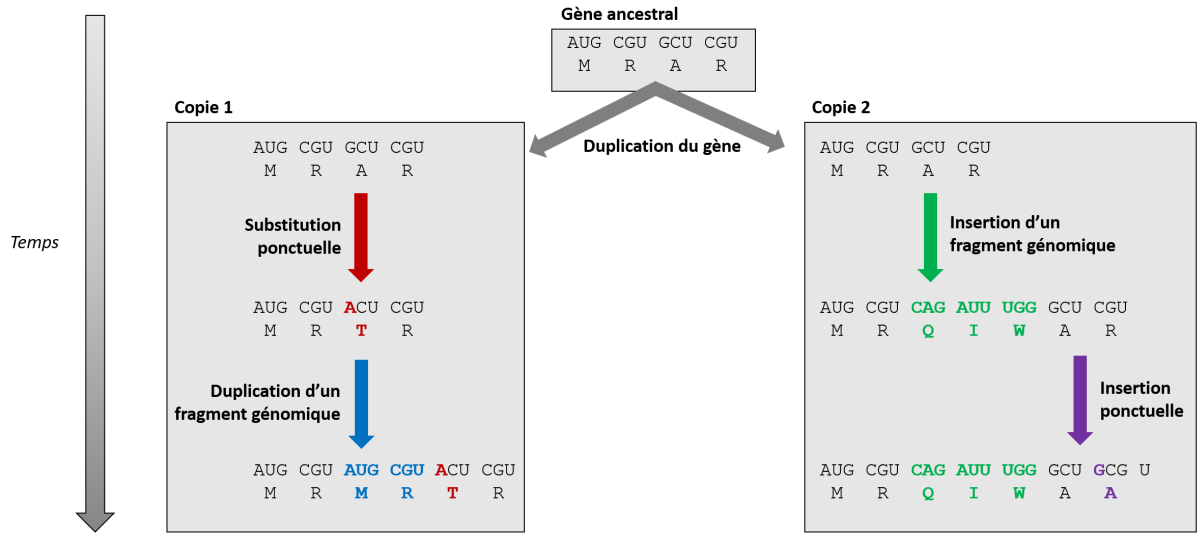

FIGURE 1.11 – **Divergence des séquences au cours du temps**
Les séquences de deux copies d'un même gène ancestral divergent indépendamment.

Les divergences de séquences peuvent également conduire à des divergences fonctionnelles (Figure 1.12). Le cas le plus parlant est celui des pseudogènes, où l'accumulation de mutations délétères sur une copie provoque une perte de fonction, jusqu'à un stade auquel le gène ne produit plus de protéine. La divergence des copies peut également entraîner la sous-fonctionnalisation (i.e., spécialisation d'une copie pour l'une des fonctions de son gène ancestral), ou la néo-fonctionnalisation (i.e., gain d'une nouvelle fonction). L'exemple des familles de protéines contenant des enzymes et des pseudoenzymes est une bonne illustration des phénomènes de divergences de séquences/fonctions [Abu+17]. Il est également possible qu'une protéine gagne des fonctions au cours de l'évolution, ce

phénomène est appelé *protein moonlighting*. L'exemple le plus étudié est celui d'enzymes qui gagnent des fonctions non catalytiques en plus de leur fonction catalytique d'origine [Jef19].

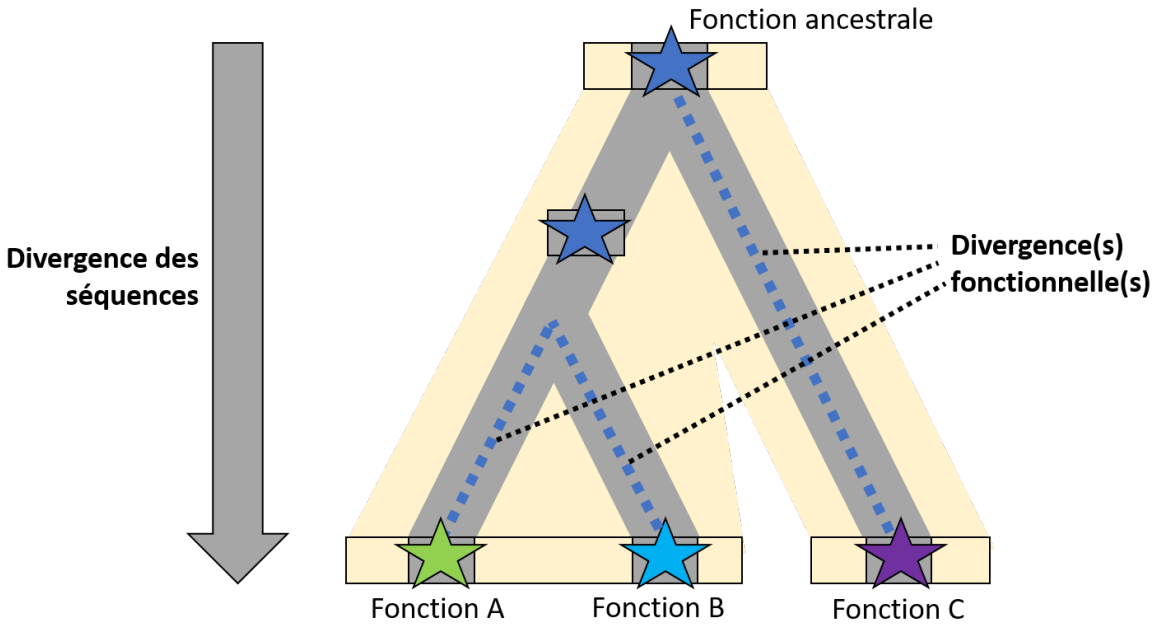

FIGURE 1.12 – **Divergence des séquences et des fonctions au cours du temps**

Ces phénomènes de divergence fonctionnelle sont généralement associés aux duplications paralogues d'un gène, de manière à expliquer la diversité fonctionnelle des copies paralogues. Alors que les copies paralogues seraient des copies aux fonctions différentes, les copies orthologues sont généralement considérées comme des copies de même fonction au sein d'espèces différentes (i.e., *The Ortholog Conjecture*) [TKL97 ; SR09]. Cependant, les divergences fonctionnelles existent également au niveau des copies orthologues, s'associant à la spécificité fonctionnelle que la protéine peut avoir au sein d'une espèce. Le critère principal d'une divergence fonctionnelle est alors la durée pendant laquelle les gènes (et leurs séquences) ont divergé depuis la duplication, plus que le type de duplication. Les copies dont la duplication est récente sont plus sujettes à partager des fonctions que des copies ayant divergées sur une plus grande durée [Sta+20].

1.3.1.3 Sélection naturelle et pressions de sélection

Les évènements de modification des séquences surviennent à différentes échelles. À l'échelle de l'individu les mutations affectent les lignées germinales, sont transmises aux descendants puis sont soumises au tri de la sélection naturelle. À l'échelle des populations une mutation est dite fixée quand l'allèle s'est substitué, au fil des générations, à tous les autres allèles. À l'échelle des espèces les différences ponctuelles fixées sont nommées substitutions pour les nucléotides et remplacement pour les aides aminés. Par commodité dans le cas des protéines, nous parlerons de substitutions dans la suite du manuscrit.

On distingue plusieurs types de pressions de sélection naturelle, neutre, positive ou négative. Lorsque les mutations ne changent pas le phénotype sous pression de sélection, ou que la pression de sélection est faible, les chances de survie de la descendance ne sont pas changées. La mutation est neutre et la sélection naturelle, dite neutre, n'a pas d'effet sur sa fixation dans la population. on parle de la dérive de la séquence.

Dans certains contextes environnementaux nécessitant des adaptations, les mutations de la séquence provoquant une modification du phénotype peuvent conférer un avantage adaptatif. Elles seront plus souvent fixées dans la population. La divergence de la séquence peut alors être plus rapide comparée à la dérive, et de nouvelles fonctions apparaissent. On parle de sélection naturelle positive. Au contraire, une fonction de la séquence peut avoir une telle importance que les mutations modifiant le phénotype sont éliminées par la sélection naturelle. La séquence diverge alors plus lentement et on parle de sélection naturelle négative.

C'est le phénomène de sélection naturelle négative qui provoque donc la conservation des séquences entre les espèces. Des régions des protéines, les domaines notamment, vont être conservées durant l'évolution car elles participent à une fonction importante. On dira par la suite que la fonction et les régions sont sous pression de sélection. Nous généraliserons ce raisonnement dans cette thèse, en interprétant les régions fortement conservées de protéines homologues comme une implication de ces régions dans une fonction ou un phénotype important.

1.3.1.4 Alignement multiple de séquences

L'alignement de séquences de protéines homologues permet d'identifier les résidus conservés au cours de l'évolution. Des protéines homologues partagent généralement une structure 3D ainsi que des fonctions similaires, héritées depuis leur ancêtre commun. Les contraintes structurelles et fonctionnelles d'une protéine appliquent une pression de sélection sur les différents résidus qui y sont impliqués, laissant des « empreintes » sur les séquences actuelles, qui peuvent alors être étudiées par la construction d'un alignement multiple de séquences (Figure 1.13, partie droite).

Un alignement multiple de séquences (MSA, du terme anglais pour *Multiple Sequence Alignement*) est une représentation d'un jeu de séquences homologues dans laquelle les résidus des différentes séquences sont alignés sur la base de leur similarité (conservation des propriétés physico-chimiques). Des caractères *gaps* sont insérés dans les différentes séquences, de manière que les séquences soient de même longueur et que les résidus supposés dériver d'une même position ancestrale puissent être alignés dans une même colonne. Pour davantage de détails sur les MSA, voir [Tho+11 ; Cha+16b ; RC20]. Un MSA permet d'identifier des résidus conservés sous la pression de la sélection naturelle.

Pazos (2022) [Paz22] distingue trois types de conservations dans un MSA (décrits sur la Figure 1.13) : 1) la conservation des résidus à une position au sein de la totalité de la famille, 2) la conservation de couplages entre résidus à plusieurs positions dues à leur coévolution et 3) la conservation des résidus à une position au sein d'une sous-famille/d'un sous-groupe de séquences.

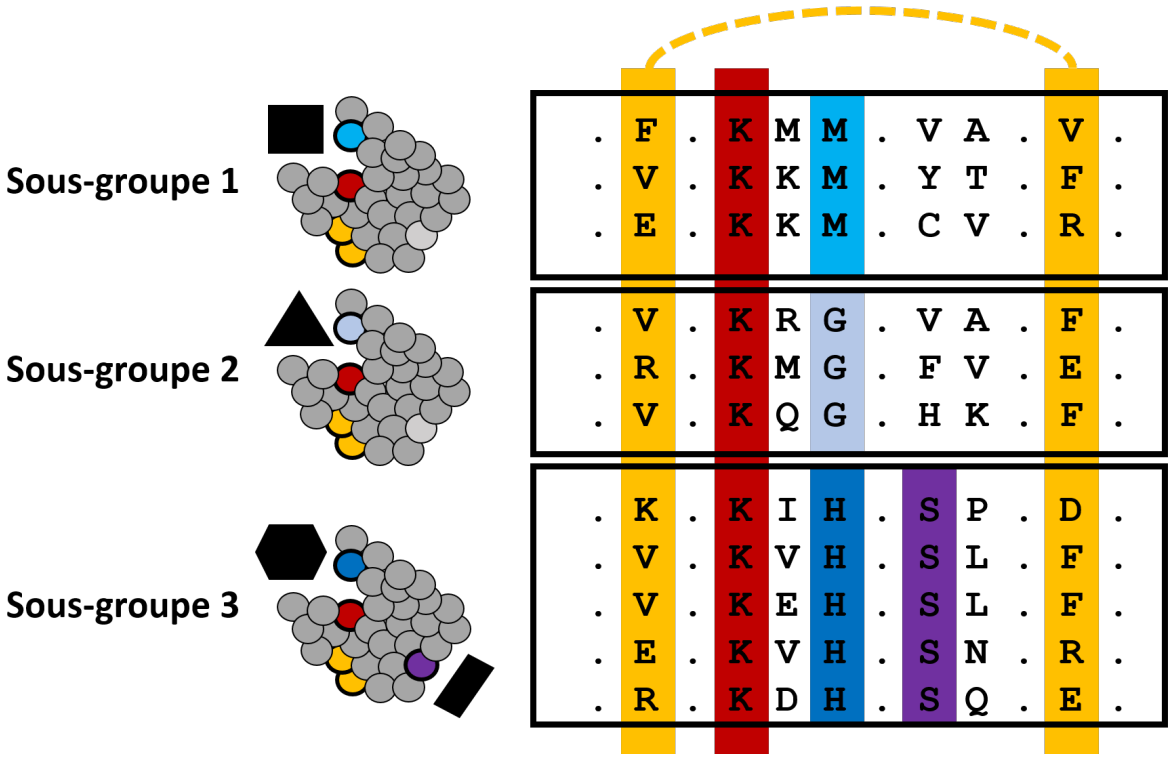

FIGURE 1.13 – **Empreinte évolutive et variations des résidus fonctionnels**, adapté de [Paz22] Représentation de l'alignement multiple de 11 séquences homologues de trois sous-groupes différents (illustrés par leurs structures communes en gris et leurs partenaires d'interactions spécifiques en noir). La colonne rouge représente une conservation d'un résidu dans la totalité des séquences homologues. La colonne en différentes teintes de bleus représente un résidu conservé différemment au sein de chaque sous-groupe, et spécifique aux partenaires de chaque sous-groupe. D'une manière similaire, la colonne en violet représente un résidu conservé uniquement au sein du sous-groupe 3. Les deux colonnes jaunes représentent deux résidus qui coévoluent.

Chapitre 2

Phylogénie moléculaire, modéliser les relations de parenté de séquences homologues

Nothing in Biology Makes Sense
Except in the Light of Evolution.

Theodosius Dobzhansky
(1964)

2.1 Reconstruction des relations de parenté de séquences homologues

Des séquences similaires peuvent être homologues et descendre d'un ancêtre commun, en reconstruire l'évolution permet une meilleure compréhension. La reconstruction de l'évolution de séquences homologues appartient au domaine de la phylogénie moléculaire, dont l'objectif principal est de déterminer l'arbre phylogénétique qui décrit au mieux les relations de parenté évolutives des séquences homologues. Les séquences homologues actuelles sont les seules observations disponibles. Elles sont issues de différents évènements évolutifs de duplications et de spéciations, ont accumulé des substitutions et ont été sujettes à des pressions de sélection. Le problème est alors de retracer l'histoire de ces processus sous forme d'un arbre phylogénétique, avec comme seule information de départ les séquences des homologues qui co-existent actuellement.

2.1.1 Arbre phylogénétique

Un arbre phylogénétique représente les liens de parenté entre des entités (e.g., espèces, séquences) et sa topologie représente leur histoire évolutive (Figure 2.1). Il est composé de nœuds et de branches, où chaque nœud correspond à une instance de l'entité observée (e.g., un gène), les entités actuelles sur lesquelles a été basée l'inférence de l'arbre sont les feuilles de l'arbre (e.g., gène actuel dont la séquence a permis l'inférence), les autres nœuds sont internes et correspondent à des nœuds ancestraux (e.g., gène ancestral). Dans le cas d'un arbre binaire, chaque nœud est lié à trois branches, et dans le cas où cet arbre est enraciné (i.e., on en connaît l'origine), une branche le lie à son ancêtre, et les deux autres branches à ses descendants. Si l'arbre n'est pas binaire, il possède au minimum un nœud qui n'est pas résolu et son nombre de descendants est alors supérieur à deux. La topologie d'un arbre peut être décrite par l'ensemble de ces bipartitions : chaque branche sépare l'arbre en deux parties, avec un sous ensemble de taxons dans chacune (e.g., `seq1 seq2 | seq3 seq4` sur la Figure 2.1). La longueur d'une banche représente la quantité d'évolution qui s'est produite entre deux nœuds (e.g., le nombre de substitutions par site). Une branche peut également être associée à des valeurs de supports statistiques (e.g., *bootstrap*).

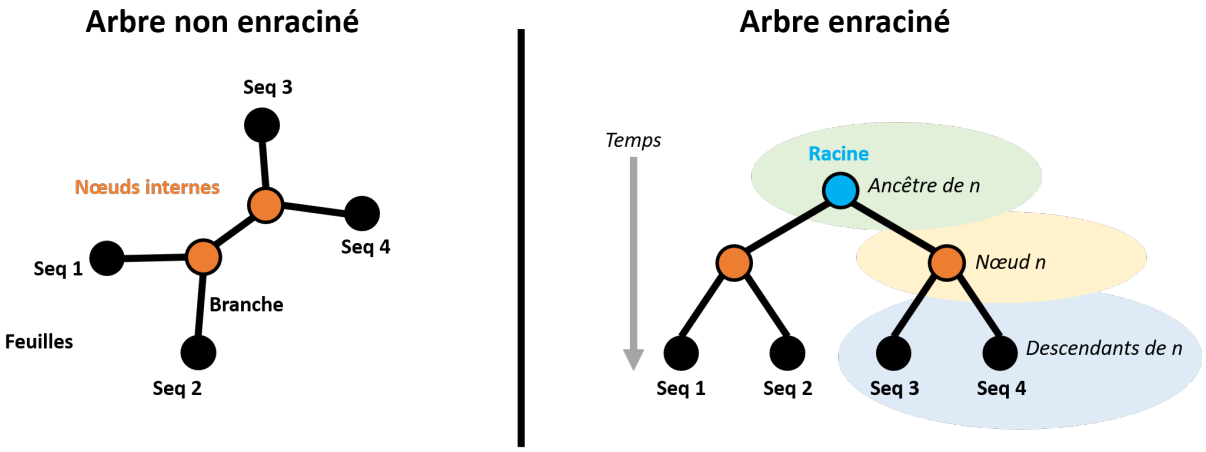

FIGURE 2.1 – **Arbre phylogénétique.**
Exemple d'un arbre phylogénétique enraciné et non enraciné représentant les liens de parenté de quatre séquences homologues : seq1, seq2, seq3 et seq4 (feuilles de l'arbre). Chaque nœud interne (correspondant à une entité ancestrale) est lié à trois autres nœuds afin de représenter les relations de parenté. Si l'arbre est enraciné (l'origine en est connue), chaque nœud n possède un ancêtre et deux descendants. La topologie de cet arbre peut être décrite par la bipartition : `seq1 seq2 | seq3 seq4`.

2.1.2 Modéliser l'évolution des séquences

L'évolution des séquences (e.g., génome, gène, virus, protéine) est généralement modélisée comme un processus stochastique (i.e., basé sur des probabilités) où chaque événement de substitution de nucléotide ou d'acide aminé au cours du temps est décrit avec une probabilité. Par exemple, pour une séquence de nucléotide, pour tout point t_n dans le temps, chaque nucléotide a une probabilité d'être substitué par un autre nucléotide parmi les quatre possibles. Une substitution est généralement décrite comme un processus markovien d'ordre 1 : la probabilité de substituer un nucléotide à t_n dépend uniquement de t_{n-1}. On parle alors de modèle de substitution markovien à temps continu. Cette modélisation représente le processus de substitution comme réversible dans le temps, la simple analyse des données moléculaires permet d'observer la distance évolutive, mais ne permet pas de savoir dans quel sens va le temps (réversibilité), ce qui a pour effet que tout arbre produit sera non enraciné. Il existe différents modèles de substitution qui permettent de modéliser les processus de transformation des séquences et ainsi d'estimer les distances évolutives entre les séquences. Ces modèles sont basés sur différents paramètres et différentes propriétés [Fin10 ; Wil+21].

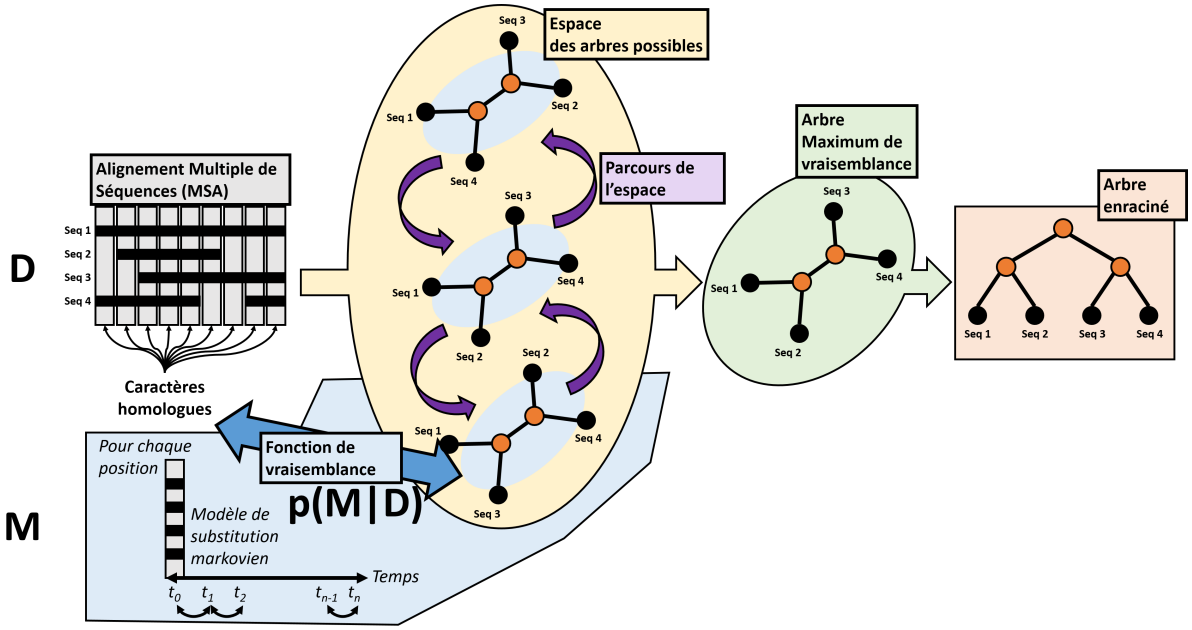

FIGURE 2.2 – **Inférence d'un arbre par phylogénie moléculaire par maximum de vraisemblance.**
Inférence d'un arbre phylogénétique (**vert** et **orange**) représentant l'histoire évolutive d'un jeu de séquences homologues (Seq1, Seq2, Seq3 et Seq4, en **noir**). Les séquences sont alignées de manière à former une matrice : l'alignement multiple de séquences (MSA, en **gris**) qui servira de base pour la construction de l'arbre. Une ligne du MSA représente une séquence, et une colonne représente des caractères homologues alignés [WM17]. Ces caractères homologues sont considérés comme hérités et dérivés depuis un ancêtre commun : des variations au cours de l'évolution y sont observées (substitutions). Si une séquence ne possède pas de caractère homologue pour une colonne, un *indel* représente l'insertion/délétion du caractère pour cette séquence. L'espace des arbres possibles regroupe toutes les topologies d'arbres possibles pour un même jeu de séquences homologues (en **jaune**). Parcourir cet espace des arbres possibles permet d'explorer les différentes topologies possibles (en **violet**) et de calculer leur vraisemblance : la probabilité des données D sachant le modèle M : p(D|M) (en **bleu**). La fonction de vraisemblance mesure à quel point les substitutions des caractères homologues observés au sein du MSA sont probables sachant le modèle d'évolution de substitution markovien. Différentes heuristiques permettent d'identifier la topologie qui maximise le score de vraisemblance dans le cas des méthodes de vraisemblances (ML pour *Maximum Likelihood*), et d'obtenir l'arbre à la topologie la plus vraisemblable (en **vert**). Cet arbre nécessite ensuite une étape supplémentaire pour être enraciné (en **orange**).

2.1.3 Construction d'un arbre phylogénétique

La première étape pour construire un arbre phylogénétique à partir de séquences homologues est de construire un alignement multiple de ces séquences (MSA, détaillé en gris sur la Figure 2.2), afin d'identifier des caractères homologues (issu d'un ancêtre commun et ayant pu subir des substitutions [WM17]) sur la base de leur similarité. Il existe une multitude de logiciels permettant de construire un MSA à partir de séquences homologues, chacun basé sur des principes différents (Figure 2.3) et étant plus ou moins efficace et adapté aux types de séquence étudiés ou à l'utilisation que l'on veut en faire [Cha+16a].

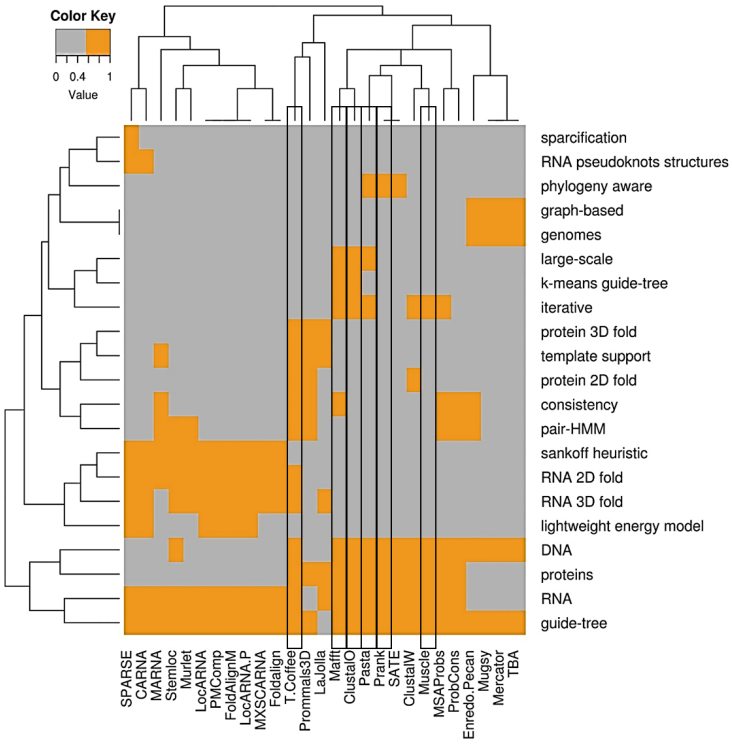

FIGURE 2.3 – **Les principaux algorithmes d'alignement multiple de séquences**, d'après [Cha+16a]
Comparaison des algorithmes de construction d'alignement multiple de séquences les plus utilisés. Les caractéristiques considérées dans une méthode sont indiquées en orange. Les dendrogrammes représentent les similarités entre aligneurs (haut de la figure) d'une part, et d'autre part les similarités entre leurs caractéristiques (à gauche). Les algorithmes qui seront utilisés dans cette thèse sont indiqués par des encadrements.

La seconde étape (qui peut être optionnelle) est le nettoyage de l'alignement, elle consiste à ne garder que les colonnes du MSA les plus informatives. Les *indels* n'étant

généralement pas considérés dans les modèles d'évolution, une colonne qui en possède une grande quantité est jugée peu informative et donc porteuse de peu de signal phylogénétique. Cependant, une colonne peu informative reste informative, le dilemme est de savoir si le nettoyage du MSA a pour effet de d'enlever du bruit ou de l'information. Les principaux programmes utilisés pour nettoyer automatiquement un MSA sont `Gblocks` [Cas00] et `Trimal` [CSG09]. De plus, le nettoyage d'un MSA peut se compléter ou se faire exclusivement manuellement.

La troisième étape est l'inférence d'un arbre phylogénétique à partir d'un MSA, nettoyé ou non (en bleu, en jaune, en violet et en vert sur la Figure 2.2), à partir de différentes heuristiques qui vont permettre d'identifier la topologie de l'arbre sachant sa « vraisemblance » : une fonction du MSA et du modèle de substitution. Les principaux programmes de reconstructions d'arbres sont présentés dans le Tableau 2.1.

La quatrième et dernière étape consiste à enraciner la topologie obtenue (en orange sur la Figure 2.2), en déterminant son origine dans le temps au milieu d'une branche, afin d'éviter toute erreur de raisonnement [Pen13]. L'enracinement d'un arbre se fait majoritairement par l'utilisation d'un groupe externe, groupe proche mais n'appartenant pas au groupe d'intérêt (un alignement doit être possible). L'enracinement de l'arbre peut également être estimé arbitrairement sur la base de la topologie de l'arbre (e.g., enraciné au milieu, enraciné en fonction de la divergence la plus récente) ou par l'utilisation d'informations extérieures (e.g., utilisation de l'arbre des espèces pour enraciner un arbre de gènes).

Dans le cas des méthodes de maximum de vraisemblance, l'arbre phylogénétique inféré est alors le plus vraisemblable, telle que la probabilité des données (MSA) soit maximale, sachant l'arbre et le modèle de substitution. Cependant, la probabilité que l'arbre soit correct (i.e., représente strictement la réalité biologique) est très faible [WM17]. Différentes méthodes peuvent alors permettre de mesurer la robustesse de l'arbre et de ses différents sous-arbres, par exemple par rééchantillonnage qui consiste à faire varier les caractères donnés en entrée (enlever des colonnes, enlever des lignes du MSA) permet d'inférer des scénarios alternatifs à partir des mêmes données [Fel85]. Itérer ce processus avec différents tirages de caractères permet alors d'observer les fréquences auxquelles les différentes bipartitions sont retrouvées. Ces fréquences indiquent pour chaque bipartition de l'arbre à quel point elle est supportée par les données (e.g., *bootstrap*).

TABLE 2.1 – **Les principaux programmes de reconstruction d'arbre par inférence phylogénétique**, adapté de [KYT20]

Programme	Approche statistique	Modèle de mixture	Parallélisé	Note	Référence
RAxML-ng	Maximum de vraisemblance	Non	Oui	Inférence d'arbre à partir d'une matrice unique ou partitionnée, conçu spécifiquement pour de très gros jeux de données	[Koz+19]
PAML	Maximum de vraisemblance	Non	Non	Comparaisons d'arbre, estimation du temps de divergence, reconstruction d'état ancestraux, simulation de séquences, détection de sélection positive	[Yan97]
PhyML	Maximum de vraisemblance	Non	Non	Inférence d'arbre à partir d'une matrice unique ou partitionnée, estimation du temps de divergence, reconstruction des états ancestraux, modèle phylogéographique (PhyloGeo) et inférence démographique (PhyREX)	[Gui+10]
FastTree	Heuristique d'un maximum de vraisemblance	Non	Oui	Conçu spécifiquement pour de très gros jeux de données (estimation de la phylogénie), nombre de modèles de substitution restreint	[PDA10]
IQ-Tree	Maximum de vraisemblance	Oui	Oui	Inférence d'arbre à partir d'une matrice unique ou partitionnée, implémente des modèles de partitions et de codons	[Min+20]
MrBayes	Inférence Bayésienne	Non	Oui	Inférence d'arbre à partir d'une matrice unique ou partitionnée, estimation du temps de divergence, état ancestraux, comparaison de topologies et modèle covarion	[HR01]
PhyloBayes	Inférence Bayésienne	Oui	Oui	Inférence d'arbre à partir d'une matrice unique ou partitionnée, estimation du temps de divergence, reconstruction des états ancestraux, simulation de séquences, analyse prédictive postérieure, validation croisée pour la comparaison de modèles, datations	[LLB09]
BEAST2	Inférence Bayésienne	Non	Oui	Conçu pour l'estimation du temps de divergence, nécessite un arbre enraciné, utilisation de Plugins	[Bou+19]
P4	Inférence Bayésienne et Maximum de vraisemblance	Non	Oui	Package Python (nécessite des connaissances du langage), très utile pour la comparaison de modèles et la simulation de séquences	[Fos04]
RevBayes	Inférence Bayésienne	Oui	Oui	En langage "Rev" (nécessite des connaissances du langage), permet d'implémenter des combinaisons de modèles pour un *framework* Bayesian	[Höh+16]

2.2 Reconstruction de scénarios ancestraux de caractères

Reconstruire l'histoire évolutive (i.e., arbre phylogénétique) de séquences ouvre différentes perspectives d'études, l'une d'elle est l'étude de l'évolution des caractères associés à ces séquences. En effet, génomes et gènes peuvent être associés à une grande variété de caractères qui, divergeant avec leurs séquences, peuvent différer entre les génomes et gènes actuels. Il est alors possible de reconstruire l'évolution de différents caractères au sein d'un arbre phylogénétique déjà établi. L'état du caractère est généralement connu pour (quasiment) toutes les feuilles de l'arbre ce qui permet, en utilisant un modèle d'évolution, d'inférer comment ce caractère a évolué au sein de l'arbre. Ce scénario est appelé « scénario ancestral » (Figure 2.4). Ces approches sont utilisées pour différents types de caractères dans le cadre de problématiques très différentes, comme les propriétés moléculaires [Wer+14 ; Bic+15 ; Bus+16 ; Che+22 ; Spe+21 ; Kwu+22], les séquences ancestrales [HLM21 ; Mas22], les phénotypes [ED09 ; Mar+12 ; BOD13 ; Sau+17 ; AKR19], les traits écologiques [Mao+17] ou les localisations géographiques [Arb15 ; Edw+11 ; Dud+17 ; GH18 ; Müh+20 ; Ari+22], c'est-à-dire tout type de caractère qui est susceptible d'évoluer et de diverger avec la séquence. D'une manière similaire à l'évolution des séquences (où un caractère est la somme des colonnes d'un MSA), l'évolution d'un caractère peut être reconstruite par une diversité d'approches probabilistes (maximum de vraisemblance [Pag99 ; RS08 ; Ish+19], méthodes Bayésiennes [HB01 ; PMB04 ; Yon+20]) ou par des approches de parcimonie [SM87], sur la base d'un arbre dont la topologie est fixée.

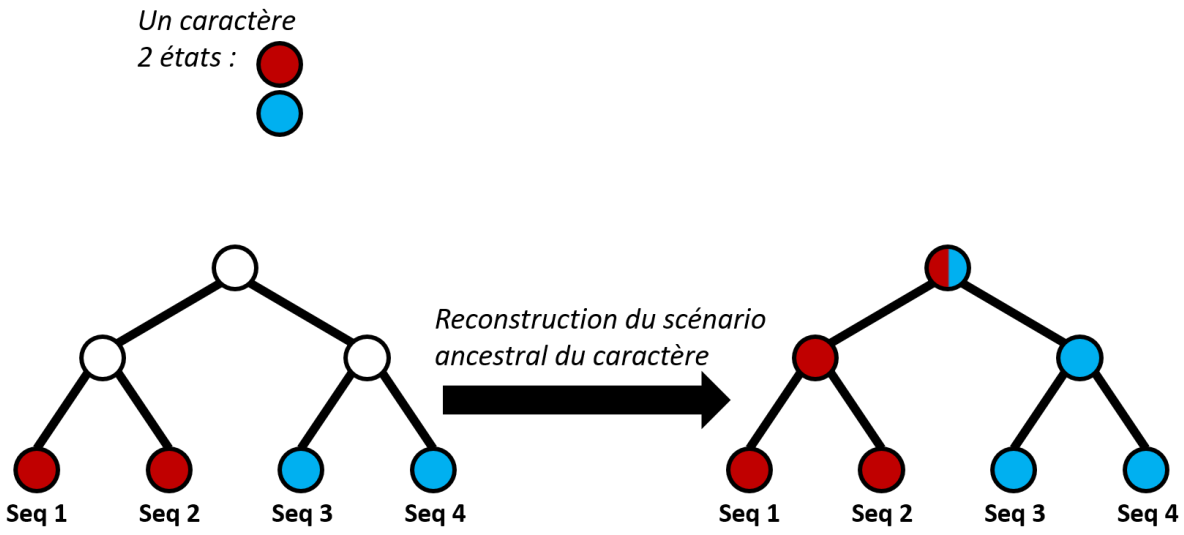

FIGURE 2.4 – **Scénario ancestral d'un caractère.**
Les feuilles de l'arbre phylogénétique de Seq1, Seq2, Seq3 et Seq4, sont annotées avec un unique caractère dont deux états sont connus (rouge et bleu). Reconstruire le scénario ancestral de ce caractère permet d'estimer son histoire évolutive au sein de l'arbre, c'est-à-dire l'état de ce caractère pour les différents nœuds internes. Si le modèle ne peut pas se prononcer, l'état du caractère peut rester ambigu, dans notre exemple, c'est le cas pour la racine à laquelle il n'est pas possible de savoir si l'état est bleu ou rouge.

2.3 Évolution d'une famille multidomaines

2.3.1 Évolution modulaire des protéines : les histoires propres des éléments d'un gène

Les familles de protéines multidomaines évoluent par brassage de domaines, un processus qui inclut l'acquisition (e.g., insertion, transfert), la duplication, la fusion, la fission et la perte de domaines [LB19b ; Oak17 ; WRK12] (Figure 2.5). Les gains, pertes et remplacements/modifications de domaines issus de ce brassage de domaines provoquent alors des changements immédiats et drastiques des fonctions de la protéine, permettant alors la variation fonctionnelle des gènes d'une même famille [Sto+15]. Le brassage des domaines résulte principalement du brassage des exons des gènes [CFS10 ; Pat96], ce dernier permet ainsi l'évolution des phénotypes et la diversité fonctionnelle des protéines d'une même famille [Kaw+09], qui inclut notamment l'émergence d'interactions protéine-protéine [CFS10]. Le brassage des exons et des domaines est un phénomène largement reconnu et étudié [SOG19], considéré comme responsable de la diversité des protéines eucaryotes, spécifiquement chez les organismes multicellulaires [GKT09] et tout particulièrement chez les protéines matricielles [Bor06 ; Pat21] (e.g., les ADAMTS-TSL).

Il est important de noter que l'évolution modulaire des protéines repose sur le brassage d'exons et n'est ainsi pas uniquement caractéristique aux domaines, mais bien à tous segments de séquences des protéines [Han+20 ; WRK12]. Le brassage d'exons et le brassage de différents segments protéiques qui en résulte a pour effet que les différents segments d'une protéine peuvent parfois évoluer de manière indépendante de leur gène. Un segment de la séquence d'une protéine peut avoir une évolution qui lui est propre en ce qui concerne au moins un évènement et qui ne dépend alors pas uniquement de l'évolution de son gène. Pour conclure, de la même manière qu'un gène peut être sujet à des évènements évolutifs indépendants de l'histoire de son génome (e.g., duplication de gènes), différents segments d'un gène peuvent être sujets à des évènements évolutifs indépendants de l'histoire de leur gène, par exemple par brassage exonique (Figure 2.5).

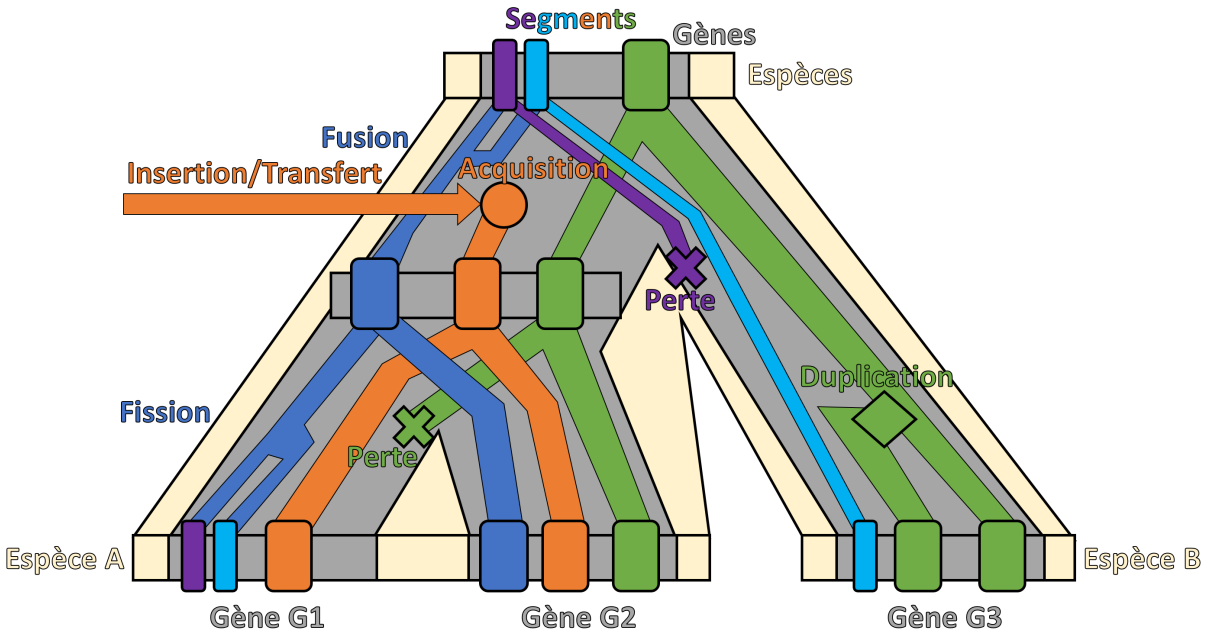

FIGURE 2.5 – **Évolution d'une famille multidomaine.**
Une famille multidomaine se distingue par trois niveaux d'évolution : **Espèces**, **Gènes** et différents **segments** de ces gènes (e.g., domaines). Les génomes contenant les gènes évoluent par spéciation et duplication. Un gène peut avoir une histoire évolutive différente de celle des espèces par duplication du gène au sein d'un génome. Ici l'arbre représente la spéciation de deux espèces (espèce A et B, arbre jaune) et la duplication du gène G dans la lignée de l'espèce A (arbre gris). Des segments de ce gène (violet, bleu foncé, bleu clair, orange, vert) peuvent également avoir des histoires évolutives qui leur sont propres, par brassage des exons de leurs gènes. Nous illustrons sur cet exemple les différents évènements d'histoires évolutives propres aux segments d'un gène. Les **segments violet et bleu clair** sont fusionnés en un unique segment qui forme alors un domaine (en bleu foncé), avant que ce segment-domaine soit hérité par un descendant et fissionné en deux segments indépendants chez un autre descendant (gène G1). Chez le gène G3, le segment violet est perdu. Le **segment orange** émerge (acquisition), par insertion ou transfert. Le **segment vert** est perdu spécifiquement chez un gène, et dupliqué chez un autre.

2.3.2 Réconciliation Domaine Gène (Espèce)

Afin de considérer les différents événements qui influent sur l'évolution d'une famille de gènes homologues, nous distinguons alors trois niveaux d'évolution associés à des événements indépendants qui leur sont propres : espèces (spéciation), gène (duplication, perte et transfert de gènes) et les différents segments d'un gène par exemple par brassage exonique. Ce dernier s'étudiant généralement à l'échelle des domaines des protéines, nous parlerons ici d'évolution des domaines. De plus, ces niveaux, et donc leurs évolutions, sont imbriqués : un domaine est porté par un gène qui est lui-même porté par un génome (espèce). Les histoires évolutives de ces trois niveaux d'évolution sont donc à la fois interdépendantes (un niveau inférieur évolue avec un niveau supérieur, subissant tous les évènements survenant au niveau supérieur) et indépendantes (événements propres à un niveau inférieur). Les histoires évolutives propres à chacun de ces niveaux peuvent être représentées par le biais de différents types d'arbres phylogénétiques : arbre des espèces, arbre de gènes et arbre de domaines. Les topologies d'arbres de niveaux différents peuvent différer [Mad97] : deux bipartitions d'arbres de niveaux différents peuvent alors être congruentes ou incongruentes, en fonction de si elles décrivent des événements évolutifs interdépendants ou indépendants.

L'histoire évolutive d'une famille de gènes peut être reconstruite par un processus de *réconciliation*, qui va comparer les topologies des différents arbres. Historiquement introduit pour expliquer les incongruences entre gènes et espèces [Goo+79], les modèles de réconciliation Gène-Espèce sont en constante évolution [Goo+79 ; NRI06 ; Nak09 ; Doy+11 ; THL11 ; Sto+12]. Il existe également différentes méthodes de réconciliation permettant de réconcilier d'autres types d'arbres phylogénétiques (e.g., Hôte-Symbiote, Géographie-Espèce), pour plus de détails sur le domaine des réconciliations phylogénétiques, voir la revue de Menet et al. (2021) [MDT21]. Nous nous intéressons ici uniquement aux réconciliations qui permettent de considérer l'évolution à l'échelle des domaines.

C'est Stolzer et al. (2015) [Sto+15] qui a proposé en premier un *framework* de réconciliation Domaine-Gène (`Notung-DM`) dans le but de reconstruire l'histoire évolutive d'une famille de protéines multidomaines considérant le brassage des domaines (i.e., possibilité d'évolution de domaine indépendante de l'histoire évolutive du gène), sur la base d'un arbre de gènes préalablement réconcilié avec l'arbre des espèces (réconciliation Gène-Espèce). L'algorithme de `Notung-DM` repose sur la comparaison des topologies d'un arbre de domaines et de l'arbre de gènes où chaque nœud de l'arbre de domaines (feuilles et nœuds internes) est associé à un nœud de l'arbre de gènes, cette association de nœuds

d'arbres différents est nommée *mapping*. Ces *mapping* sont alors expliqués par les évènements suivants : co-divergence (le domaine diverge avec le gène, de manière synchrone), duplication, perte, insertion et transfert horizontal de domaine. Chaque évènement est associé à une valeur, ce qui permet de rechercher la réconciliation la plus parcimonieuse parmi toutes celles possibles. Une réconciliation possible est une réconciliation temporellement faisable, où chaque transfert qui y est proposé implique des espèces qui auraient coexisté. Le *framework* `Notung-DM` nécessite un arbre de gènes enraciné et binaire, un arbre de domaines également enraciné et binaire, la correspondance (i.e., *mapping*) entre toutes les feuilles de l'arbre de domaines et les feuilles de l'arbre de gènes (i.e., gènes dont provient chaque domaine). Il permet alors d'obtenir la réconciliation Domaine-Gène la plus parcimonieuse et temporellement faisable : cette dernière présente les *mapping* associant nœuds ancestraux de l'arbre de gènes et nœuds ancestraux de l'arbre des espèces, ainsi que les évènements associés.

2.3.3 Réconciliation Domaine Gène Espèce avec SEADOG

Bien que l'approche de réconciliation de Stolzer (Gène-Espèce puis Domaine-Gène) soit considérée comme avant-gardiste, elle ne modélise pas l'interdépendance entre les évolutions des domaines, gènes et espèces. C'est cette limite qui a motivé Bansal et al., (2018) [LB19b ; LB18] à proposer une modèle de réconciliation *Domain-Gene-Species* (DGS) dans le but de réconcilier conjointement l'évolution des domaines, des gènes et des espèces.

La réconciliation DGS est implémentée dans le programme `SEADOG` [LB19b ; LB18] qui prend en entrée un arbre des espèces, un arbre de gènes (ou plusieurs) et un arbre de domaines dont les feuilles sont formatées de manière à indiquer de quel gène provient un domaine et de quelle espèce provient un gène (pour un gène : `NomGène_NomEspèce` ; pour un domaine : `NomDomaine_NomGène_NomArbreGène`). La réconciliation DGS modélise l'évolution des domaines, gènes et espèces par le biais de deux types de réconciliation : Gène-Espèce et Domaine-Gène. Chacun des nœuds internes de l'arbre d'un domaine est associé à un nœud interne de l'arbre des gènes (*mapping* Domaine-Gène) et chacun des nœuds internes de l'arbre des gènes est associé à un nœud interne de l'arbre des espèces (*mapping* Gène-Espèce). La réconciliation Gène-Espèce est décrite avec des évènements de spéciation, de duplication et de perte de gènes. La réconciliation Domaine-Gène est elle décrite par des événements de co-divergence, transfert de domaine, duplication de domaine et perte de domaine. La réconciliation Gène-Espèce dépend de la réconciliation Domaine-Gène et réciproquement cette dernière dépend de la réconciliation Gène-Espèce.

C'est l'optimisation conjointe des scores de leurs évènements qui permet de proposer la réconciliation la plus parcimonieuse. À noter que ce modèle n'impose pas de contraintes temporelles aux évènements de transferts, l'inconsistance temporelle n'ayant que très peu d'impact sur la précision d'une réconciliation [Ban+15]. Le résultat d'une réconciliation DGS avec `SEADOG` comprend (Figure 2.6) :

- **Tous les arbres** qui ont été réconciliés avec les différents nœuds internes labellisés par un indice unique (`DX`, `GX` et `SX` pour respectivement les domaines, les gènes, les espèces), suivi des indices de ses descendants (`GX_G1_G2`, `G1` et `G2` étant les descendants du gène ancestral `GX`).
- **Les *mapping*** qui associent les nœuds des différents d'arbres : Gène-Espèce et Domaine-Espèce, ainsi que les évènements qui y sont associés.
- **Le coût total** de la réconciliation et un résumé des événements inférés.

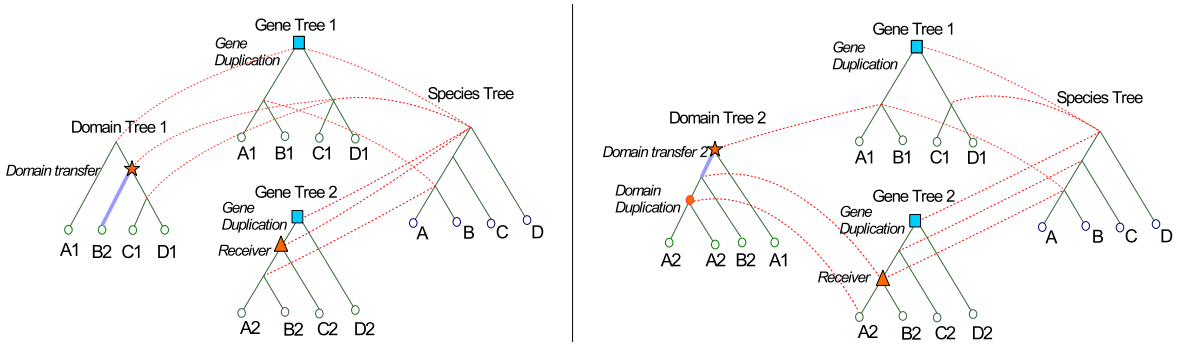

FIGURE 2.6 – **Réconciliation Domaine-Gène-Espèce**, d'après [LB19c]
Deux réconciliations DGS indépendantes : un arbre de domaine est réconcilié avec deux arbres de gènes et un arbre des espèces. Les *mapping* entre les arbres sont représentés par des pointillés rouges. La réconciliation Gène-Espèce étant différente entre les deux réconciliations DGS, le *mapping* du gène C2 l'est également (c'est le seul nœud à différer).

Cependant, considérer et optimiser conjointement les réconciliations Domaine-Gène et Gène-Espèce signifie que les réconciliations pour deux domaines différents peuvent produire deux scénarios Gène-Espèce différents non compatibles entre eux (le nœud C2 est réconcilié différemment sur la Figure 2.6). La réconciliation Gène-Espèce est alors optimisée indépendamment pour chaque domaine. C'est pourquoi Bansal et al. (2019) [LB19c] propose avec `SEADOG-MD` une réconciliation mDGS : multiDomaine-Gène-Espèce, qui permet de considérer simultanément tous les domaines, en introduisant pour cela la notion de *Groupe-Domaines/Gène* (association des différents domaines actuels à leur gène actuel). La réconciliation mDGS optimise d'une manière conjointe la réconciliation Gène-Espèce et

toutes les réconciliations Domaine-Gènes (Figure 2.7). Les entrées, sorties et évènements considérés par `SEADOG-MD` sont les mêmes que `SEADOG`. Tout comme `SEADOG`, `SEADOG-MD` s'utilise intégralement en ligne de commande et ne fournit pour l'instant pas de visualisations ou de sorties au format standard de description de réconciliation RecPhyloXML (2018) [Duc+18].

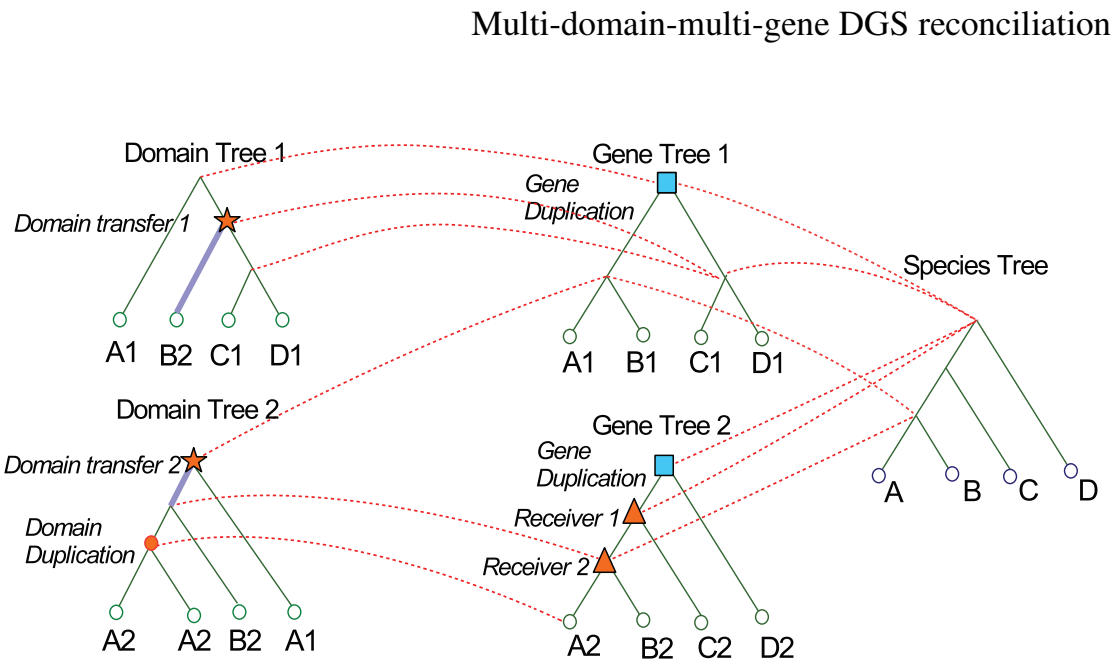

FIGURE 2.7 – **Réconciliation multiDomaine-Gène-Espèce**, d'après [LB19c]
Réconciliation mDGS de deux arbres de domaines, deux arbres de gènes (1 et 2) et un arbre des espèces (A, B, C, D). Les *mapping* entre les arbres sont représentés par des pointillés rouges.

Conclusion

Reconstruire l'histoire évolutive de séquences homologues consiste à construire un arbre phylogénétique qui va représenter leurs relations de parenté. Un arbre phylogénétique peut alors être utilisé pour étudier l'évolution et la divergence de divers caractères des séquences homologues. Dans le cas d'une famille de protéines homologues multidomaines, trois niveaux d'évolution (espèces, gènes et segments géniques) sont interdépendants tout en possédant des évènements qui leur sont propres. L'évolution de chaque niveau peut être représentée par un arbre phylogénétique et ces différents arbres peuvent être réconciliés afin de modéliser l'évolution de la famille.

Chapitre 3

Annotation fonctionnelle de protéines

It may be said that natural selection is daily and hourly scrutinizing, throughout the world, the slightest variations ; rejecting those that are bad, preserving or adding up all that are good ; silently and insensibly working, whenever and wherever opportunity offers ... We see nothing of these slow changes in progress, until the hand of time has marked the lapse of ages, and then so imperfect is our view into long-past geological ages, that we see only that the forms of life are now different from what they formerly were.

Charles Darwin
L'Origine des espèces (1859)

Une fonction, ou « annotation fonctionnelle » d'une protéine, est considérée dans ce manuscrit comme tous types de connaissance biologique pouvant être associés à une séquence de protéine (e.g., localisation cellulaire, implication dans un processus biologique, capacité à interagir avec des partenaires spécifiques). Le nombre croissant de séquences de protéines disponibles n'est pas corrélé à une croissance équivalente des connaissances biologiques associées. Il est estimé que moins de 1 % des protéines présentes dans la base de données UniProtKB/TrEMBL sont caractérisées expérimentalement [Bou+16]. Cette réalité motive le développement de méthodes de prédictions des fonctions des protéines

[Rau+21 ; Sma+21]. Les méthodes de prédiction peuvent exploiter différents types de données expérimentales : les séquences, les structures et les données hauts débits (e.g., PPI, expression des gènes). Les différents types d'information utilisés par les méthodes de prédiction sont présentés en Figure 3.1. Les séquences sont les informations les plus largement disponibles. Dans sa revue des méthodes de prédiction, Shehu et al., (2016) [SBM16] classe les méthodes basées sur les séquences en trois catégories : les méthodes par homologie, les méthodes phylogénomiques et les méthodes par contexte génomique. Dans ce chapitre, nous allons présenter les méthodes qui reposent uniquement les séquences de protéine, c'est-à-dire les méthodes par homologie (Section 3.1) et les méthodes phylogénomiques (Section 3.2), l'approche que nous proposons dans cette thèse étant à l'intersection des deux.

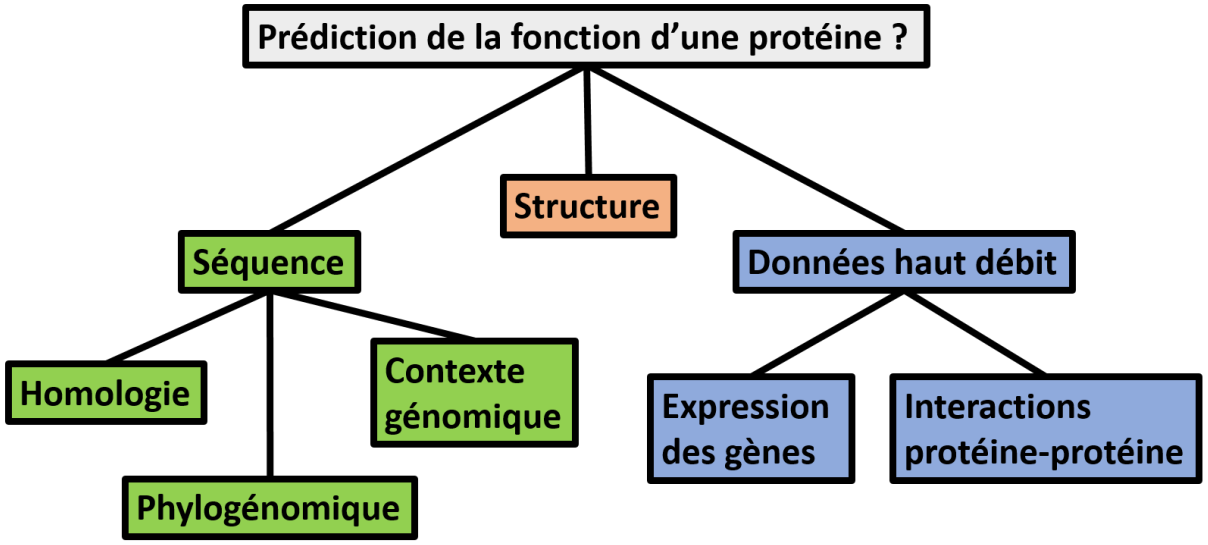

FIGURE 3.1 – **Principales informations utilisées par les méthodes de prédictions fonctionnelles des protéines**.

3.1 Propagation d'annotations par homologie

Une annotation fonctionnelle peut être propagée à une protéine non annotée sur la base de la similarité de sa séquence avec celle d'une protéine déjà annotée, à condition que sa séquence partage une similarité significative. Une approche standard est d'utiliser l'outil d'alignement `BLAST` [Alt+90] pour aligner la séquence non annotée avec des séquences

provenant de bases de données de séquences, de manière à identifier des protéines aux séquences significativement similaires qui seraient déjà annotées. Cette annotation est alors propagée à la séquence non annotée. L'hypothèse sous-jacente est que si deux séquences partagent un fort degré de similarité, donc des propriétés physicochimiques proches, elles ont évolué depuis un ancêtre commun et possèdent alors des fonctions similaires, voire identiques.

Différentes méthodes permettent de ne pas se limiter à considérer une unique séquence annotée, et d'utiliser à la place une modélisation d'un ensemble de séquences partageant une même annotation. De telles approches permettent d'utiliser des *signatures* plus sensibles qu'une unique séquence. Ces dernières donnent plus d'importance aux résidus importants pour la fonction/structure de la protéine (qui sont alors conservés) [Fri06], et moins (ou pas) d'importance aux résidus moins importants (qui sont alors peu ou pas conservés). Par exemple, la méthode PSI-BLAST [Alt+97] va permettre de rechercher des homologues lointains, par un processus itératif de recherche de séquences similaires, dont l'ensemble des séquences trouvées est modélisé par un *profile* (décrit en Section 3.1.2). Il est alors possible d'interroger à nouveaux les bases de données de séquences, mais cette fois-ci à partir du *profile*. Les séquences trouvées seront moins spécifiques à la séquence initiale, mais partagent les mêmes résidus conservés.

Il existe différents types de signatures qui permettent de modéliser l'information caractéristique d'un regroupement de séquences sur la base de l'identification de conservation (e.g., motifs, PSSM, pHMM). En fonction du critère de regroupement des protéines qu'elles modélisent (e.g., partage d'une fonction, partage d'un même domaine), les signatures peuvent décrire différents types de séquences biologiques (i.e., superfamille et famille de protéines homologues, domaines, sites spécifiques, régions désordonnées). Une fois établies, les signatures sont généralement stockées dans différentes bases de données. Le consortium `InterPro` [Blu+20] (Figure 3.2) intègre 13 bases de données de signatures de protéines (ainsi que la base `MobiDB` pour les régions désordonnées), permettant l'utilisation des signatures préétablies pour annoter de nouvelles séquences de protéines et la propagation d'annotations.

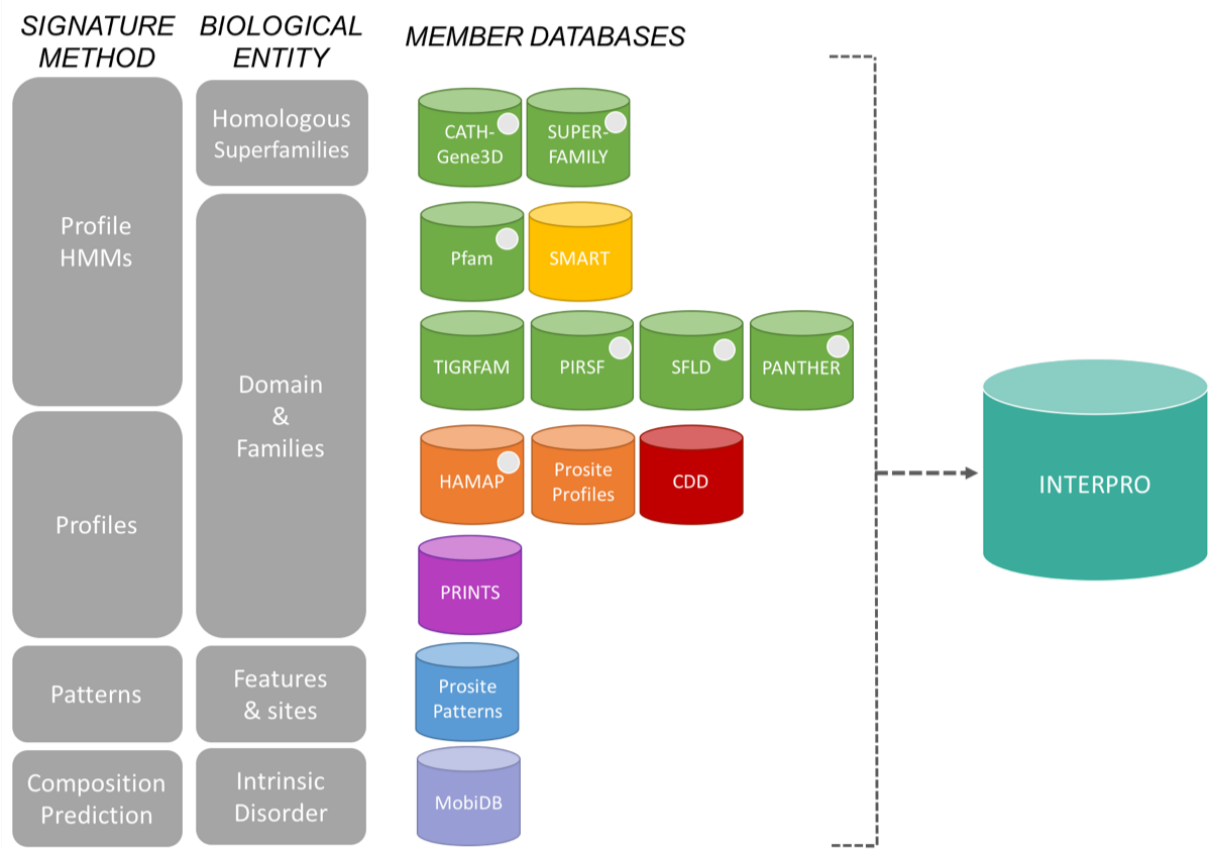

FIGURE 3.2 – **Le consortium InterPro**, d'après [ELI]

3.1.1 Modéliser les conservations d'un MSA

L'approche standard pour modéliser un ensemble de séquences repose sur la construction d'un alignement multiple (MSA) qui permet d'identifier les positions (colonne du MSA) où des résidus sont conservés. Les colonnes conservées d'un MSA peuvent être représentées simplement par un motif. Un motif est une expression régulière qui représente les variations possibles aux différentes positions ainsi que les distances entre celles-ci (exemple en Figure 3.3). Les motifs Prosite [Sig+02] sont des exemples de motifs comme signatures.

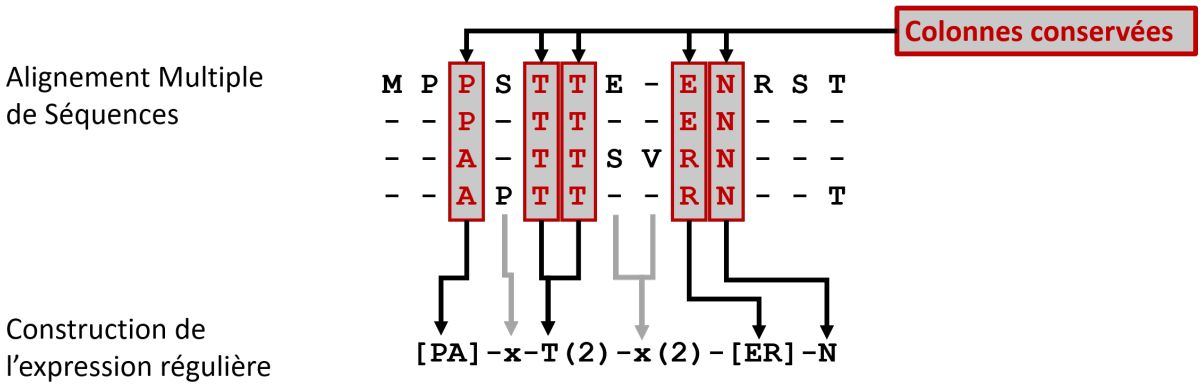

FIGURE 3.3 – **Un motif comme représentation de résidus conservés.**
Les colonnes conservées d'un alignement multiple de séquences homologues (MSA) peuvent être représentées par un motif (e.g., motif Prosite), sous la forme d'une expression régulière. L'expression régulière présentée ici se lit de la manière suivante : [Pro ou Ala] - n'importe quel acide aminé - Thr - Thr - n'importe quel acide aminé - n'importe quel acide aminé - [Glu ou Arg] - Asn.

Il est également possible de modéliser un MSA dans sa globalité via un profil HMM [Kro+94]. Un profil HMM (i.e., pHMM *profile Hidden Markov models*) est une représentation probabiliste d'un MSA, qui modélise sur la base d'un modèle statistique les conservations, les insertions et les délétions (exemple en Figure 3.4). À chaque position conservée du MSA, correspondent trois états : un état de *match* (la position est conservée chez la majorité des séquences homologues), un état d'insertion (permet d'ajouter des positions entre deux *match*) et un état de délétion (qui permet de sauter un *match*). Les probabilités de transition entre les états sont estimées sur la base du MSA. Les domaines Pfam [Mis+21] sont par exemple modélisés par des profils HMM.

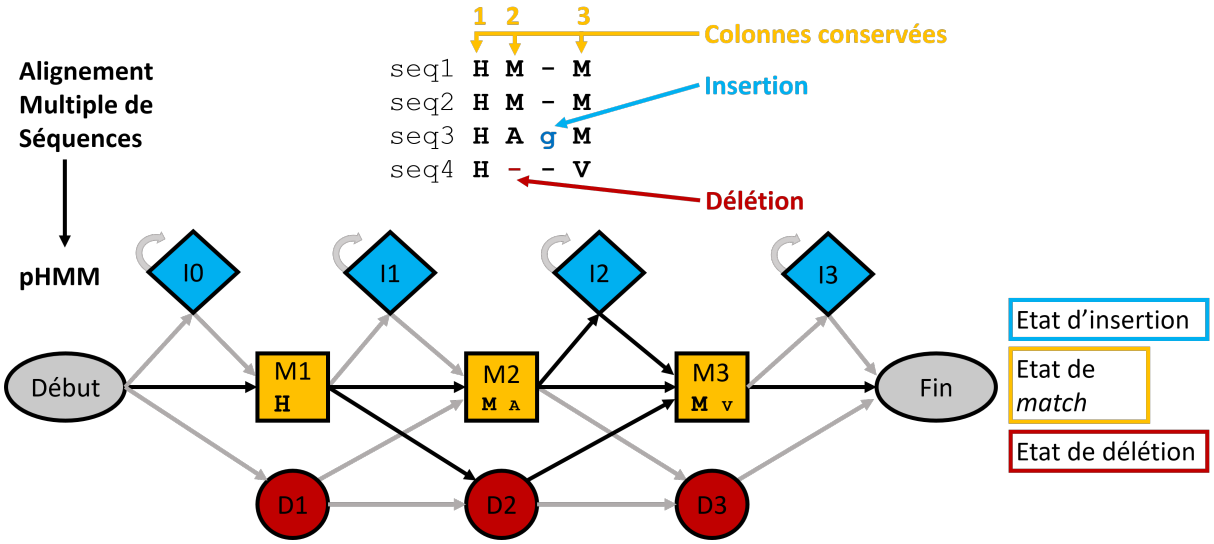

FIGURE 3.4 – **Un profil HMM comme représentation d'un alignement multiple de séquences.**
Représentation d'un alignement multiple de séquences (MSA) en profil HMM (pHMM). Les probabilités des résidus conservés sont modélisées dans les états de *match* (en jaune). Les probabilités d'enchaînements des résidus, d'insertions et de délétions sont modélisées par les flèches entre les états de *match* (jaune), d'insertion (bleu) et de délétion (rouge).

3.1.2 Modéliser une région conservée d'un MSA

Une PSSM [Sch+86] (*Position-Specific Scoring System*) permet de modéliser par une matrice une région localement conservée (Figure 3.5). Une PSSM est une matrice qui recense le (poids) de chaque acide aminé à chaque position, généralement le log du rapport de vraisemblance entre la fréquence observée et attendue des acides aminés à la position. Les PSSM ne permettant pas de considérer les insertions et les délétions, les *profiles* [GME87] en sont une extension qui le permet.

Afin de représenter plusieurs régions conservations d'une même famille, les *fingerprints* [Att+96] modélisent une signature par un enchaînement ordonné de PSSM (ou de *profiles*) et une indication de la distance les séparant, ce qui permet alors de considérer les conservations spécifiques à des groupes de séquences.

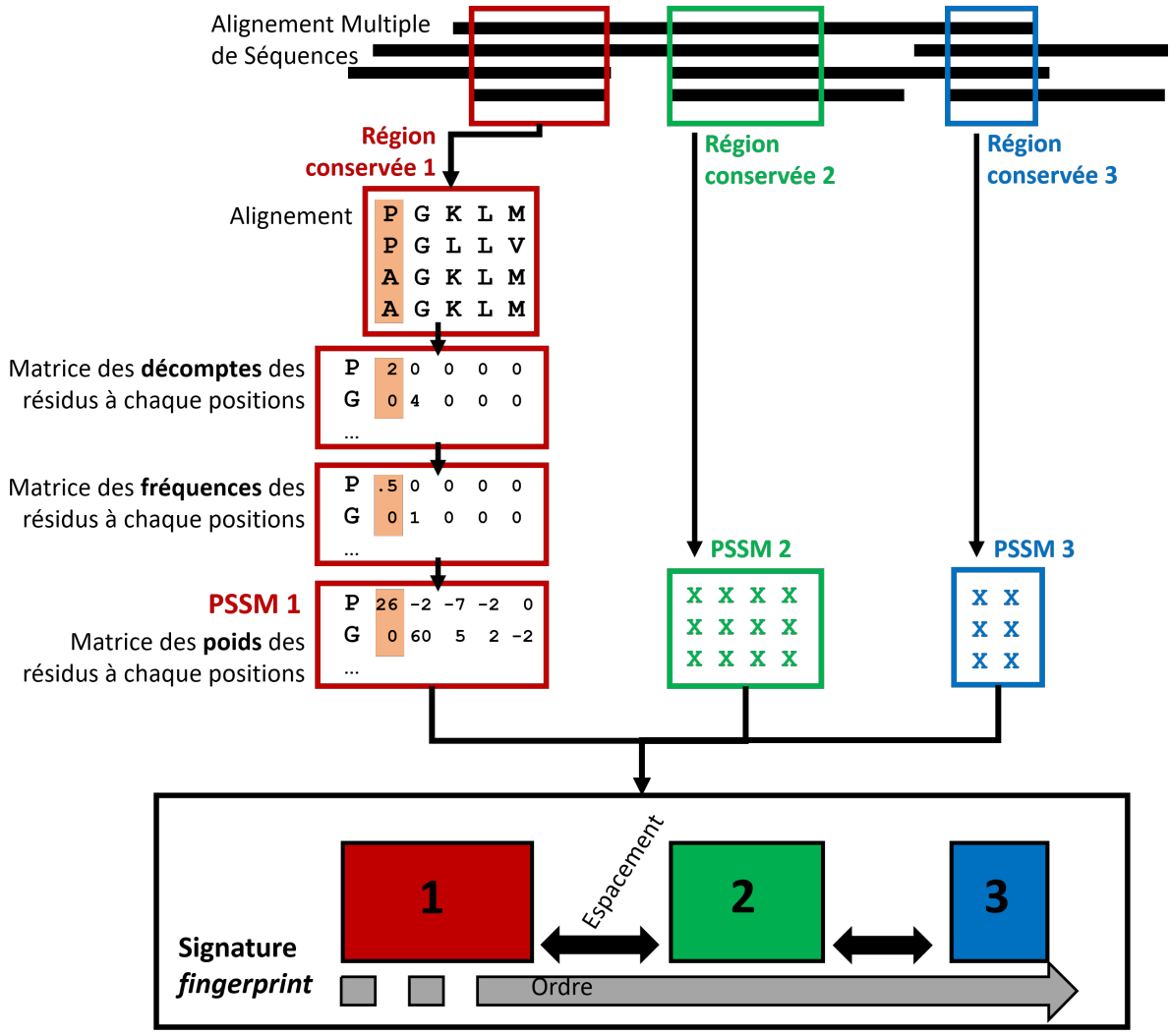

FIGURE 3.5 – **PSSM et Fingerprint comme représentations de régions conservées.**
Une région d'un alignement multiple de séquences homologues (MSA) est représentée sous forme d'une PSSM, qui correspond à une matrice comportant tous les poids de position d'acides aminés de l'alignement. Par exemple, pour le motif 1, la position colorée en orange possède dans sa matrice un poids pour chacun des 20 acides aminés existants. Le regroupement de PSSM de différentes régions localement conservées d'un même alignement multiple permet de constituer une signature *fingerprint* de la famille de protéines homologues, qui contient les PSSM ordonnées et leur espacement.

3.1.3 Modéliser l'enchaînement possible des régions conservées via un PLMA

La suite `Protomata` [Ker08 ; Cos22] a introduit en 2008 une méthode de modélisation d'un jeu de séquences, par construction d'un automate des séquences appelé *protomate* (Figure 3.6). Un protomate peut être décrit comme les différents enchaînements possibles de PSSM. La particularité de cette méthode qui nous intéresse ici est que la construction d'un protomate s'affranchit des limitations inhérentes aux méthodes de modélisations basées sur l'utilisation d'un alignement multiple (MSA), par introduction d'un nouveau type d'alignement : l'alignement multiple partiel local (PLMA, pour *Partial Local Multiple Alignment*).

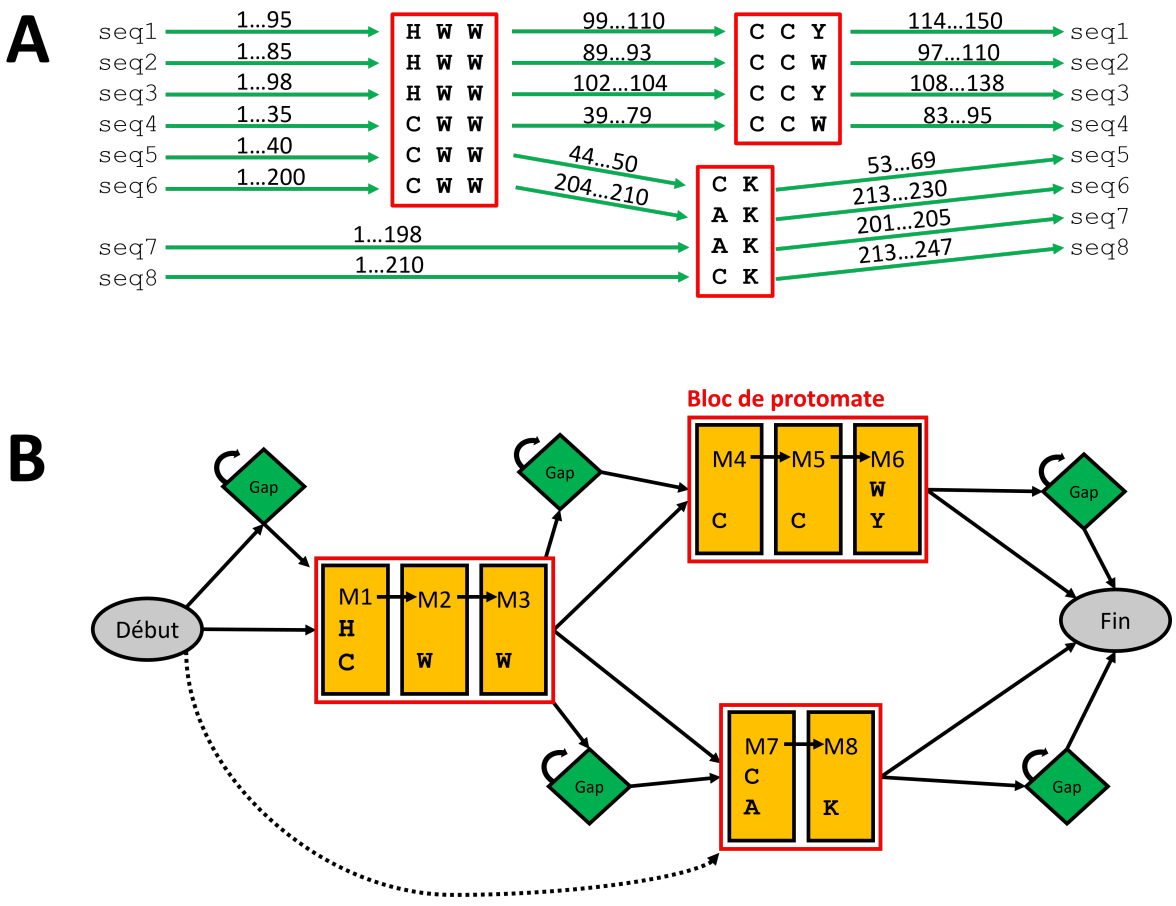

FIGURE 3.6 – **Protomate et alignement multiple partiel local.**
Alignement multiple partiel local (PLMA) de huit séquences **(A)** et protomate construit depuis ce PLMA **(B)**.

Le programme `paloma` de la suite `Protomata` permet de construire un PLMA à partir d'un ensemble d'alignements locaux de paires de séquences de manière à identifier directement les régions localement conservées par des séquences homologues (i.e., blocs de conservation). Un PLMA permet d'identifier l'enchaînement de différents blocs de conservations locales, caractéristiques d'une famille de protéines homologues, tout en considérant l'hétérogénéité des différentes séquences homologues (Figure 3.7). Un bloc est un alignement local impliquant un sous-groupe de séquences, qui peut contenir au minimum deux séquences et au maximum toutes les séquences homologues à aligner.

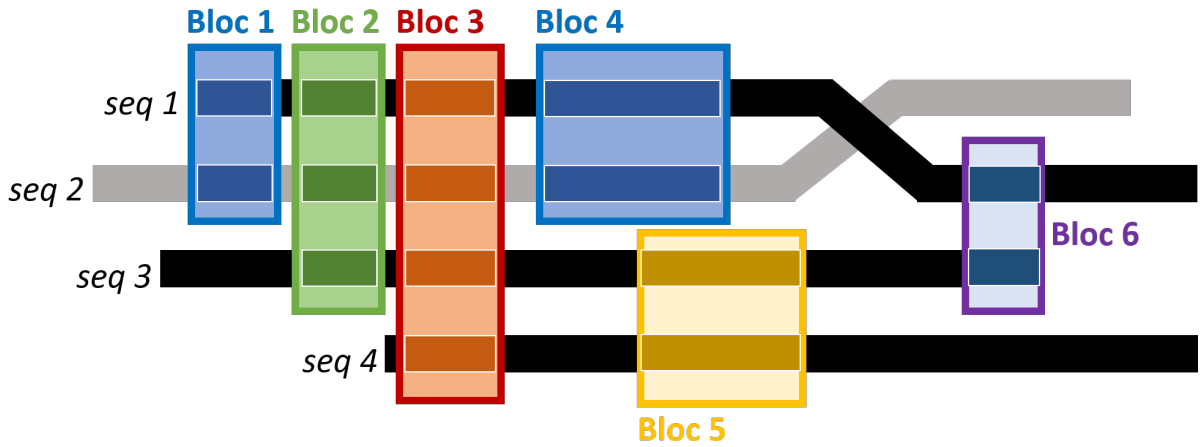

FIGURE 3.7 – **Blocs Paloma de différents sous-groupes.**
Les quatre séquences seq1, seq2 (en gris), seq3 et seq4 sont segmentées en six blocs, conservés par différents sous-groupes de séquences. Le bloc 3 est conservé au sein de toutes les séquences (en rouge). Le bloc 2 est conservé au sein des séquences du sous-groupe seq1, seq2 et seq3 (en vert). Les blocs 1 et 4 sont conservés au sein des séquences du sous-groupe seq1, seq2 (en bleu). Le bloc 5 est conservé au sein des séquences du sous-groupe seq3, seq4 (en jaune). Le bloc 6 est conservé au sein des séquences du sous-groupe seq1, seq3.

La construction d'un PLMA par `paloma` comporte trois grandes étapes (Figure 3.8). Dans un premier temps, tous les alignements locaux possibles ne contenant pas d'*indels* sont calculés avec le programme `dialign2-2` [Mor99] (i.e., diagonales des séquences à aligner). Dans un second temps, toutes les diagonales en accord avec les paramètres choisis sont sélectionnées. Pour finir, ces diagonales sélectionnées sont intégrées de manière itérative par alignement transitif des positions, sous contrainte de conserver la consistance globale de l'alignement [Kec93]. Le PLMA obtenu est alors composé de blocs conservés, ne contenant aucun *gap*. Un bloc contient un ensemble de segments d'au minimum deux séquences (mais pas forcément toutes les séquences). Au sein d'un bloc, chaque position

d'un segment est alignée avec une position des autres segments du bloc, et n'est pas alignée avec une autre position du reste de l'alignement PLMA.

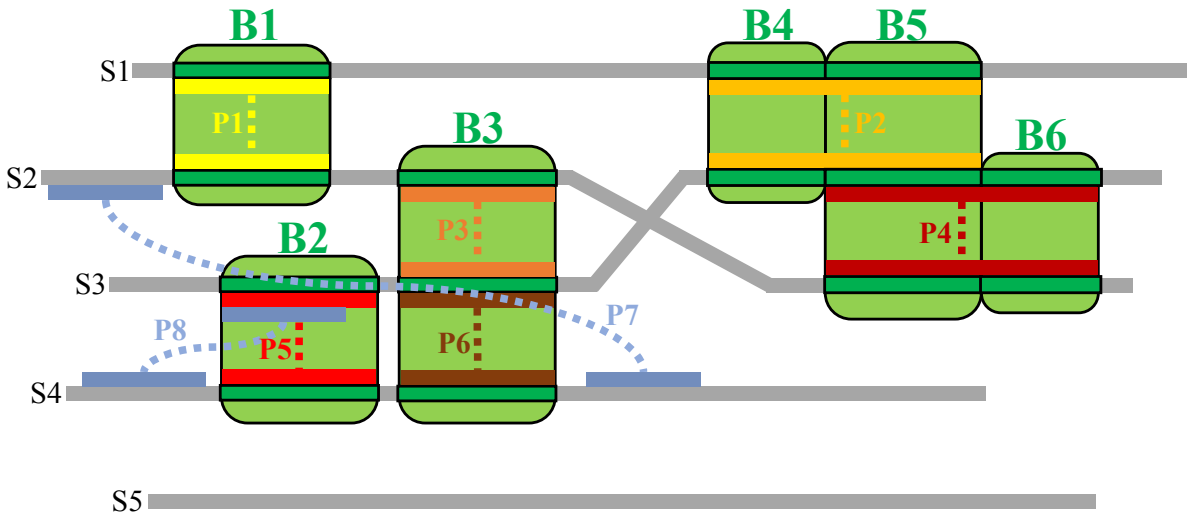

FIGURE 3.8 – **Alignement Multiple Partiel et Local par Paloma.**
Exemple schématique du PLMA des séquences S1...S5. P1...P6 sont des alignements locaux de paires de segments sélectionnés par `paloma`. P7 et P8 sont des alignements non sélectionnés, inconsistants avec P1...P6. Les blocs B1...B6 sont construits par `paloma`. Un bloc est défini comme l'alignement local maximal impliquant un sous-ensemble de séquences, inclus dans la clôture transitive $\mathcal{A}$ des alignements locaux par paires P1...P6, de manière que chaque position soit alignée localement au sein des séquences du sous-ensemble, sans être alignée à une autre séquence $\mathcal{A}$. Par exemple, la clôture transitive de P2 et P4 décrit trois blocs d'alignements : B4 pour les positions contiguës entre S1 et S3, B5 pour celles entre S1, S2 et S3 et B6 pour celles entre S2 et S3. Un bloc peut alors être plus court que l'alignement local minimal choisi. Ou plus long, par extension d'alignements locaux contigus qui se chevauchent pour la même paire de séquences (par exemple P1 qui est le résultat de la concaténation des alignements locaux S1 et S2).

3.2 Méthode de prédiction fonctionnelle phylogénomique

Les méthodes phylogénomiques proposent de considérer les histoires évolutives entre les organismes de manière à identifier des similarités fonctionnelles entre les gènes. Deux grandes approches se distinguent : l'approche par profil et l'approche par arbre.

3.2.1 Prédictions de PPI par profilage phylogénétique

La génomique comparative permet de prédire des PPI par profilage phylogénétique (Figure 3.9) [Pel+99 ; DM15] : des protéines possédant un même profil d'orthologues chez des espèces actuelles partagent une même histoire évolutive, ce qui suggère l'existence d'une interaction. La PPI exerce une pression de sélection qui a pour effet que la perte d'une protéine est suivie de la perte de son partenaire d'interaction.

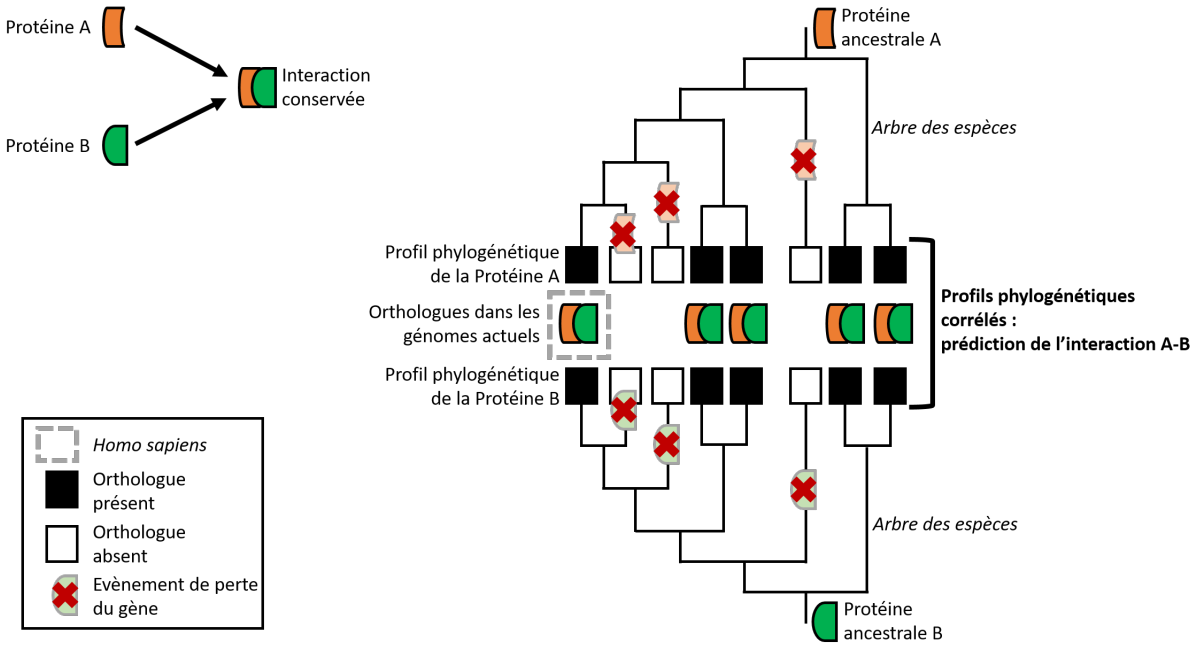

FIGURE 3.9 – **Profilage phylogénétique**, adapté de [DM15]
Illustration de la méthode de profilage phylogénétique. La présence des orthologues de deux protéines (A et B, en orange et vert respectivement) est étudiée au sein de 8 espèces. Une protéine est décrite par un profil phylogénétique qui correspond à la présence/absence d'orthologue chez les différentes espèces. Les protéines A et B ont des profils phylogénétiques corrélés, ce qui témoigne d'une contrainte évolutive (i.e., quand une protéine est perdue, la deuxième l'est également), ce qui permet de prédire l'interaction entre la protéine A et la protéine B.

3.2.2 Propager des fonctions sur la base d'un arbre phylogénétique

Proposée par Eisen [Eis98], l'approche phylogénomique standard (également nommée approche par « arbre » [SBM16]) utilise la phylogénie moléculaire dans le cadre de l'analyse fonctionnelle des protéines, et permet de ne pas utiliser uniquement les similarités de séquences, mais de considérer également leur histoire évolutive. La phylogénomique repose sur l'hypothèse que la fonction d'une protéine évolue en parallèle de sa séquence. Cette hypothèse implique qu'une phylogénie des protéines homologues va représenter la manière dont la fonction a évolué au sein de ces protéines homologues [Eng+05], permettant ainsi la propagation d'une fonction des gènes annotés aux gènes non annotés par le biais de leur arbre phylogénétique (Figure 3.10). Engelhardt et al., (2005) [Eng+05] propose d'identifier manuellement les nœuds de duplication des gènes (en contraste aux nœuds de spéciation), afin de propager les fonctions aux gènes issus d'une même duplication (i.e., orthologues). La méthode de Gaudet et al., (2011) [Gau+11] nécessite également une estimation manuelle du nœud ancestral de l'arbre dans lequel la fonction est acquise, avant de pouvoir propager la fonction à ses descendants. Yon et al., (2020) [Yon+20] proposent une approche automatisée, basée sur une méthode bayésienne de reconstruction de scénarios ancestraux (voir Section 2.2), dans le but de propager les fonctions des feuilles annotées, aux nœuds internes et aux feuilles non annotées.

À noter que, l'approche phylogénomique et les approches standards de reconstructions de scénarios ancestraux propagent des annotations au sein d'un arbre phylogénétique, mais diffèrent selon le type de nœud auquel elles cherchent à prédire un état. L'approche phylogénomique cherche à estimer l'état (i.e., la présence de l'annotation) des feuilles à partir d'un nombre limité d'observations (un nombre réduit de feuilles sont annotées). La reconstruction de scénarios ancestraux cherche généralement à estimer l'état des nœuds internes, voire de la racine, à partir de feuilles qui possèdent toutes une observation (e.g., étude d'une épidémie virale où la provenance géographique de chaque feuille est connue).

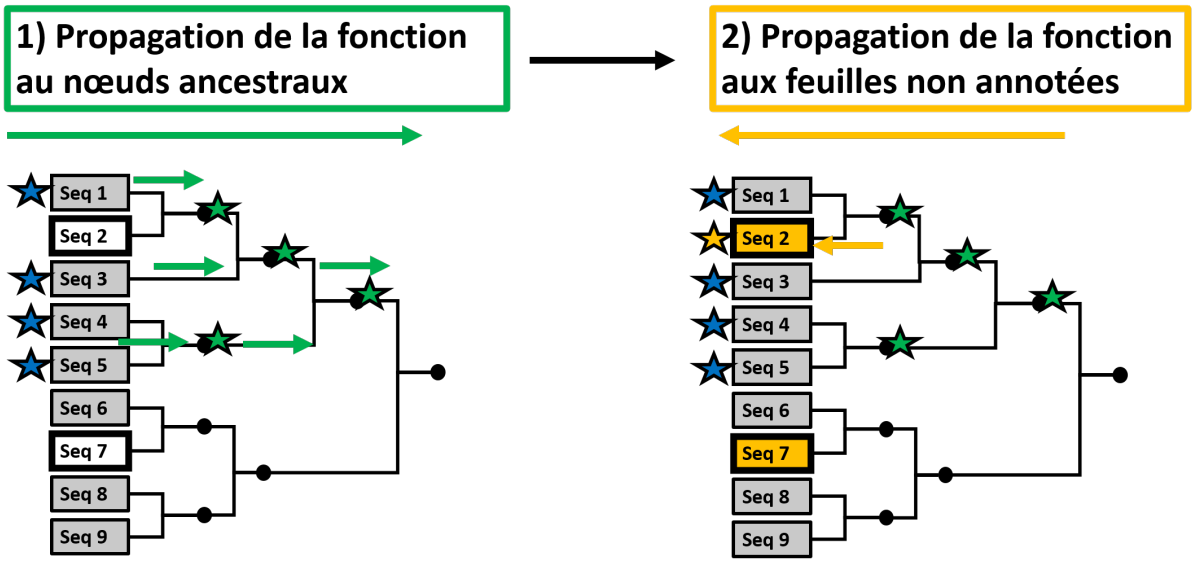

FIGURE 3.10 – **Approche phylogénomique**
Propagation de la fonction par approche phylogénomique. **1)** La fonction représentée par une étoile bleue est propagée aux nœuds internes d'un arbre phylogénétique comprenant les séquences annotées et les séquences non annotées (Seq2 et Seq7). La propagation aux nœuds internes est représentée en vert. **2)** La fonction est ensuite propagée aux feuilles non annotées, ici à la séquence Seq 2. La propagation aux feuilles est représentée en jaune.

Chapitre 4

ADAMTS-TSL, une famille multigène, multidomaine et multifonctionelle

Les protéines ADAMTS (***A** **D**isintegrin-like **A**nd **M**etalloproteinase with **T**hrombospondin motifs*) et ADAMTSL (ADAMTS-***L**ike*) sont des protéines sécrétées impliquées dans le remodelage de la matrice extracellulaire et associées à de nombreuses pathologies chez l'humain (e.g., cancer, fibrose, arthrite et maladies cardio-vasculaires) [Kel+15 ; MA18 ; HA15 ; Cai+16].

Les ADAMTS sont des métalloprotéases de la superfamille des Metzincins. Les metzincins sont caractérisées par un domaine protéase à zinc [Hux+07] et se divisent en quatre familles : les Astacins, les Matrixins, les Serralysines et les Adamalysines (Figure 4.1). Les Adamalysines comprennent les ADAMTS ainsi que les ADAM (***A** **D**isintegrin-like **A**nd **M**etalloproteinase*) et les SVMP (***S**nake **V**enom **P**roteins*) de classe III. Les ADAM sont pour la quasi-totalité des protéases transmembranaires, dont certaines protéines isoformes sont sécrétées (e.g., ADAM9 et ADAM12)[Thé+21], alors que les SVMP sont des protéases sécrétées spécifiquement dans le venin de serpent [Kal+19]. La superfamille des metzincins comprend 80 gènes humains, dont les protéines sont impliquées dans le remodelage de la matrice extracellulaire et inhibées par les TIMP (*Tissue Inhibitors of MetalloProteinases*) [Hux+07 ; BW21a ; Riv+10].

Chez l'humain, 26 protéines sont décrites comme appartenant à la famille des ADAMTS-TSL (19 ADAMTS et 7 ADAMTSL). Elles partagent avec les 20 ADAM une organisation multidomaine (Figure 4.2, [Thé+21]) qui inclue : un signal-peptide, un pro-peptide, un domaine métallopeptidase, un domaine disintégrine et un domaine riche en cystéines. Les ADAM se singularisent par un domaine EGF-like (*Epidermal Growth Factor*) et des domaines transmembranaire et cytoplasmique. Les ADAMTS sont caractérisées par une région ancillaire qui comprend un motif thrombospondine de type 1 (TSP1), un do-

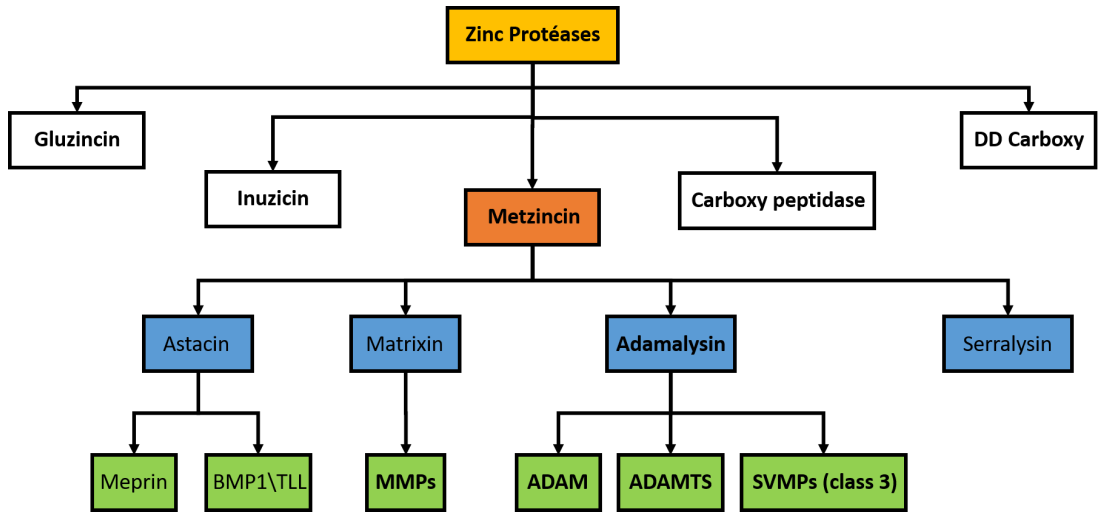

FIGURE 4.1 – **Superfamille des Metzincins**, adapté de [Hux+07].
Représentation schématique des sous-divisions de la superfamille des metzincins, au sein des Zincs protéases. Les ADAMTS sont des zincs protéases, des metzincins et des adamalysines. Les ADAMTSL par leur non-activité catalytique ne peuvent pas être considérées comme des zincs protéases et ne sont donc pas représentées.

maine spacer et des domaines/motifs additionnels incluant notamment des motifs TSP1. Contrairement aux ADAM et aux ADAMTS, les ADAMTSL sont des protéines sécrétées qui ne possèdent pas de domaine catalytique et de domaine disintégrine. Les ADAMTSL ne partagent avec les ADAMTS que la « région ancillaire ».

Les protéines ADAMTS sont classifiées en quatre sous-groupes sur la base de leurs substrats [Ros+21] : **1) les procollagénases** (ADAMTS-2, -3, -4), **2) la protéase clivant le facteur de Von Willebrand** (ADAMTS-13) et **3) les hyalectanases** (ADAMTS-1, -4, -5, -8, -9, -15, -20). **4) Les autres ADAMTS** (ADAMTS-6, -7, -10, -12, -16, -17, -18, -19) ont longtemps été considérées comme orphelines, cependant des études récentes ont permis d'identifier des substrats. Ce quatrième sous-groupe se caractérise aussi par des associations au réseau de fibrilline/fibronectine. Contrairement aux ADAMTS, les ADAMTSL (ADAMTSL-1, -2, -3, -4, -5, -6, et Papilin) ont fait l'objet de peu d'études et sont principalement caractérisées par leurs rôles de stabilisation des réseaux de microfibrilles et leur implication dans la biodisponibilité du facteur de croissance TGF-β [Le +08 ; Le +11 ; Tsu+10].

Dans ce chapitre, nous allons nous focaliser sur l'évolution des gènes ADAMTS-TSL (Section 4.1), sur la combinatoire en domaines des protéines ADAMTS-TSL (Section 4.2), ainsi que sur leur diversité et multiplicité fonctionnelle (Section 4.3).

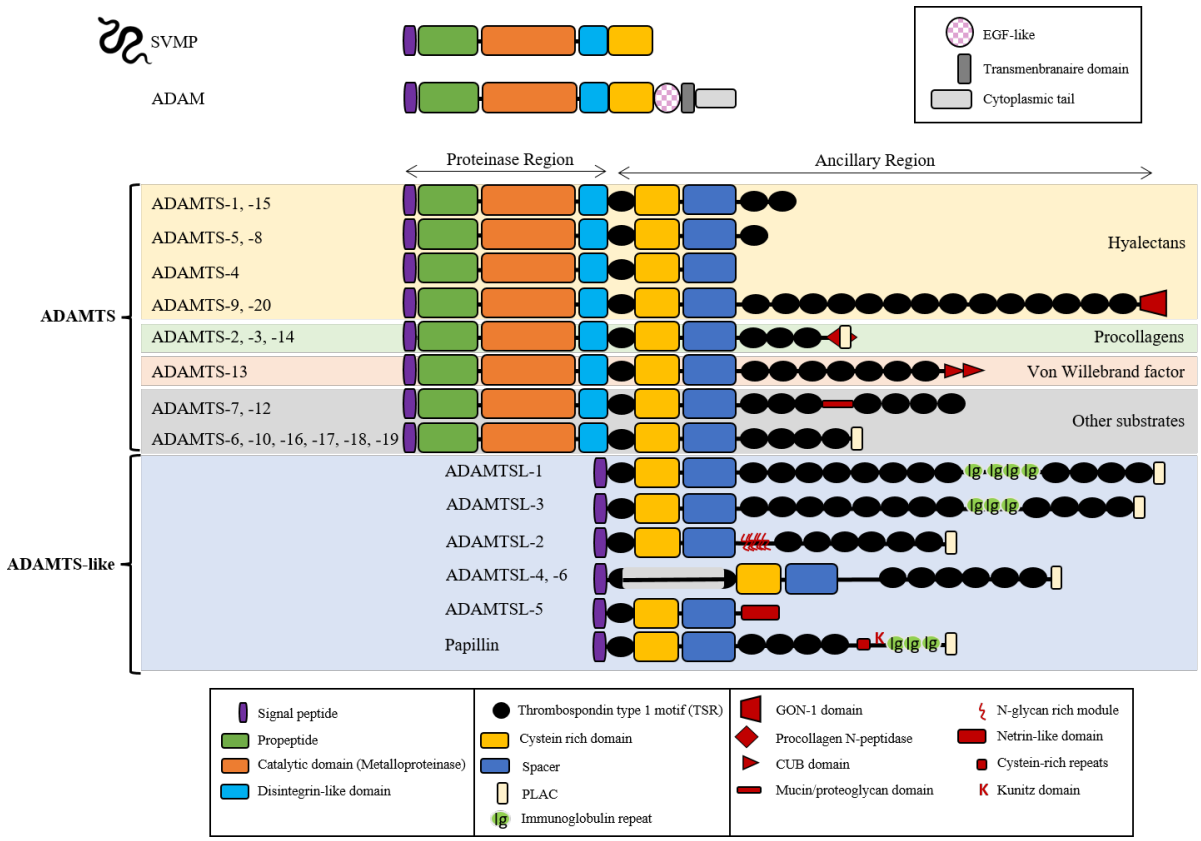

FIGURE 4.2 – **Composition en domaines et motifs des protéines ADAMTS-TSL humaines**, adapté de [Thé+21].
Les protéines partageant une même composition en domaines/motifs sont regroupées sur une même ligne. Les quatre sous-groupes caractérisés sur la base des substrats communs sont indiqués par les couleurs jaune/vert/rouge/gris en arrière-plan, les ADAMTSL étant indiquées par un arrière-plan de couleur bleu. La composition en domaines des ADAM humaines et des SVMP est également représentée.

4.1 L'évolution d'une famille multigène

Différentes études proposent des scénarios pour expliquer l'histoire évolutive des ADAMTS-TSL. Nous synthétisons ici les phylogénies (Figure 4.3) que ces études proposent et leurs conclusions cherchant à expliquer évolutivement la présence du grand nombre de copies paralogues chez l'humain.

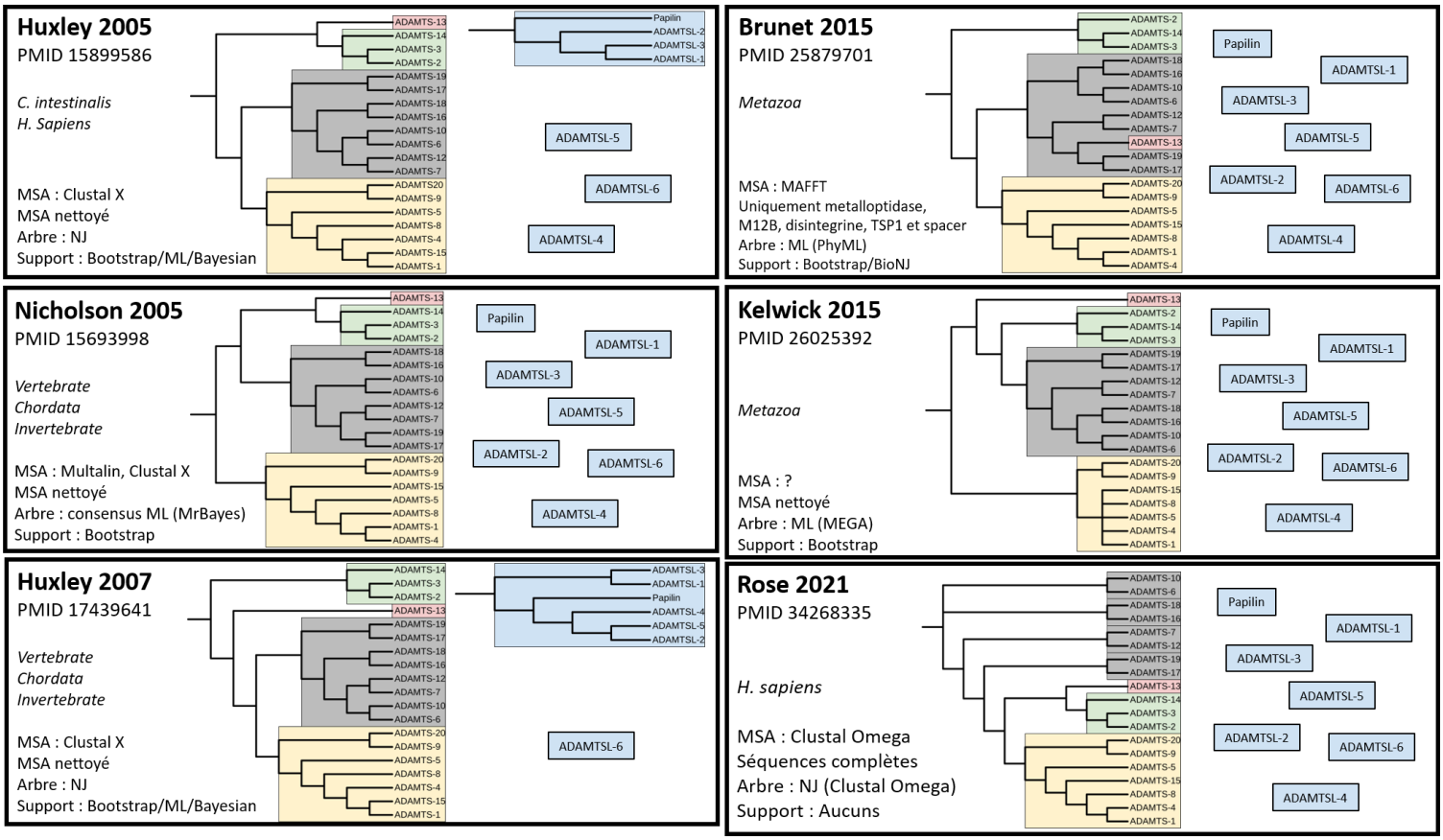

FIGURE 4.3 – **Phylogénies des ADAMTS-TSL proposées dans la littérature.**
Synthèse des phylogénies proposées par les six études proposant des arbres décrivant l'histoire évolutive des ADAMTS-TSL : Huxley et al. (2005) [Hux+05], Nicholson et al. (2005) [Nic+05], Huxley et al. (2007) [Hux+07], Brunet et al. (2015) [Bru+15], Kelwick et al. (2015) [Kel+15] et Rose et al. (2021) [Ros+21]. Seuls les 26 gènes humains sont représentés et les gènes non considérés au sein d'une étude sont représentés seuls. Les procollagénases sont colorés en vert, le facteur de Von Willebrand en rouge, les hyalectanases en jaune, les autres ADAMTS en gris et les ADAMTSL en bleu. MSA : *Multiple Sequences Alignement*, NJ : *Neighbor Joining*, ML : *Maximum Likelihood*, (?) : information indisponible.

4.1.1 Des duplications chez les vertébrés

La découverte de six gènes ADAMTS chez *Ciona intestinalis* a motivé Huxley-Jones et al. (2005) [Hux+05] à étudier l'évolution des ADAMTS et de certaines ADAMTSL (ADAMTSL-1, -2, -3 et Papilin) chez les vertébrés. Pour ceci, ils ont utilisé les ADAMTS et les ADAMTSL de trois espèces (*Homo sapiens, Ciona intestinalis, Drosophila melanogaster*) afin d'étudier l'évolution des gènes homologues. Ils proposent deux phylogénies distinctes, une pour les ADAMTS, une pour les ADAMTSL. Leur phylogénie des ADAMTS offre une explication au regroupement par sous-groupes fonctionnels (par substrat) qui était préalablement supposé, et leur phylogénie des ADAMTSL suggère que la Papilin et les autres ADAMTSL auraient divergé avant l'ancêtre *Chordata*. Nicholson et al. (2005) [Nic+05] propose également un scénario pour l'histoire évolutive des ADAMTS, cette fois en utilisant des espèces de vertébrés (*Homo sapiens, Mus musculus, Fugu rubripes*), une espèce de chordé (*Ciona intestinalis*) et des espèces invertébrées (*Caenorhabditis elegans, Drosophila melanogaster*). Leur étude est basée sur l'utilisation de méthodes phylogénétiques et d'analyses des structures exoniques des gènes et confirme les origines évolutives des sous-groupes fonctionnels, tout en proposant un scénario des différentes duplications de gènes au cours de l'évolution. À la suite de leur première étude, Huxley-Jones et al. (2007) [Hux+07] proposent également des phylogénies des différentes familles de la superfamille des metzincin, comprenant les ADAM, ADAMTS, BMPI/TLL, meprin et MMP. Leur conclusion est que la complexité de toutes ces familles de metzincin chez les vertébrés actuels a largement été acquise durant l'évolution des vertébrés.

4.1.2 Les événements évolutifs à l'origine de la complexité de la famille

Plus récemment, Brunet et al. (2015) [Bru+15] ont utilisé un grand nombre d'espèces de vertébrés afin de proposer un scénario des événements clés de l'évolution des ADAMTS au sein des vertébrés. Le scénario propose l'existence d'un seul gène ancestral ADAMTS à l'origine de la multicellularité et l'embryogenèse, dupliqué en six gènes ancestraux chez l'ancêtre des *Bilateria* (7/12 - 6/10 - 17/19 - 16/18 - 2/3/14/13 - 9/20). Le gène ancestral 9/20 aurait alors donné naissance par retrotransposition au gène ancestral 1/4/5/8/15, et le gène ancestral 2/3/14/13 aurait été dupliqué en deux gènes : 1/2/14 et 13. Ces huit gènes ancestraux des *Deuterostomia* (7/12 - 6/10 - 17/19 - 16/18 - 2/3/14 - 13 - 9/20 - 1/4/5/8/15) produiront les 19 gènes ADAMTS humains par différents événements de

duplication, notamment des génomes.

En conclusion, ces six études proposent six scénarios différents de l'évolution des gènes ADAMTS qui s'accordent sur une monophylie des sous-groupes procollagénases et hyalectanases. Elles ne sont cependant pas en accord sur la monophylie des gènes ADAMTS associés au réseau de fibrilline/fibronectine, mais les regroupent tous par paires : ADAMTS-6/-10, ADAMTS-7/-12, ADAMTS-16/-18 et ADAMTS-17/-19. Seules les études d'Huxley (2005 et 2007) ont étudié l'évolution des gènes ADAMTSL, indépendamment des ADAMTS et sans tous les considérer. Finalement, toutes ces études sont en désaccord sur les relations ancestrales entre les différents sous-groupes.

4.2 Des protéines multidomaines

Les 26 protéines ADAMTS-TSL humaines sont généralement décrites par leur composition en domaines/motifs (Figure 4.2). Les séquences des ADAMTS se décomposent en deux régions : 1) une région protéinases (incluant un peptide signal, un propeptide, un domaine catalytique et un domaine distintégrine) et 2) une région ancillaire (incluant un domaine riche en cystéines, un domaine spacer, un nombre variable de motifs thrombospondines TSP1 et divers autres domaines/motifs). Les protéines ADAMTS-TSL sont en général définies par une composition en domaines qui leur est propre, mais certaines protéines possèdent strictement la même composition. C'est le cas de ADAMTS-1 et de ADAMTS-15 ou de ADAMTSL-4 et de ADAMTSL-6, suggérant qu'elles sont issues de duplications récentes des gènes [MA18]. Les seuls domaines/motifs retrouvés chez toutes les séquences des protéines ADAMTS-TSL humaines sont : le motif TSP1 central, le domaine riche en cystéine et le spacer. Ces domaines correspondent à la partie N-terminale de la région ancillaire, le reste de la région ancillaire étant variable d'une protéine ou d'un groupe de protéines à l'autre.

Seules 4 des 26 protéines ADAMTS-TSL humaines possèdent des structures disponibles obtenues expérimentalement (Tableau 4.1) : ADAMTS-1, -4, -5, (hyalectanases) et ADAMTS-13 qui clive le facteur de Von Willebrand. Ce sont également les ADAMTS les plus étudiées. Cela signifie qu'il n'existe aucune information expérimentale pour les structures des autres sous-groupes ADAMTS (procollagénases et fibrillines/fibronectines associées) et des ADAMTSL. Et pour les structures existantes, elles ne sont pas complètes et contiennent principalement la région centrale des protéines avec le domaine catalytique. Les extrémités N-term et C-term sont absentes des structures et sont prédites comme

désordonnées.

TABLE 4.1 – **Structures expérimentales disponibles.**

Protéine	Nb structures	Méthode	Résolutions	Taille séquence	Résidus structures
ADAMTS-1	4	X-ray	2.10-2.33Å	967	253-548
ADAMTS-2	0			1211	
ADAMTS-3	0			1211	
ADAMTS-4	5	X-ray	1.24-2.80Å	837	213-520
ADAMTS-5	7	X-ray	1.40-2.60Å	930	262-628
ADAMTS-6	0			1117	
ADAMTS-7	0			1686	
ADAMTS-8	0			889	
ADAMTS-9	0			1935	
ADAMTS-10	0			1103	
ADAMTS-12	0			1594	
ADAMTS-13	5	X-ray	2.60-2.80Å	1427	79-685,1185-1427
ADAMTS-14	0			1223	
ADAMTS-15	0			950	
ADAMTS-16	0			1224	
ADAMTS-17	0			1095	
ADAMTS-18	0			1221	
ADAMTS-19	0			1207	
ADAMTS-20	0			1910	
Papilin	0			1278	
ADAMTSL-1	0			1762	
ADAMTSL-2	0			951	
ADAMTSL-3	0			1691	
ADAMTSL-4	0			1074	
ADAMTSL-5	0			481	
ADAMTSL-6	0			1018	

L'outil de prédiction `AlphaFold` [Jum+21 ; Var+22] permet de prédire des structures pour les ADAMTS-TSL qui n'en possédaient pas expérimentalement, ainsi que des structures complètes pour celles dont seul le domaine catalytique était cristallisé. Les compositions en domaines et les structures prédites par `AlphaFold` et disponibles depuis la base de données Uniprot [BA99] [1] pour les ADAMTS-TSL humaines sont présentées par sous-

1. Ces structures prédites incluent notamment les extrémités N-ter (peptides signaux, propeptides) qui peuvent être absentes des protéines matures.

groupe : les 7 hyalectanases (Figure 4.4), ADAMTS-13 (Figure 4.5), les 4 procollagénases (Figure 4.6), les 8 ADAMTS associées aux fibrillines-fibronectines (Figure 4.7) et les 7 ADAMTS-like (Figure 4.8). Les résidus y sont colorés par un gradient qui représente la confiance des prédictions : de bleu pour les résidus dont les prédictions sont de bonne qualité, à orange pour les résidus dont les prédictions sont de mauvaise qualité. Les structures prédites des 26 ADAMTS-TSL présentent toutes un cœur central structuré dont les prédictions sont de bonne qualité (en bleu), qui correspond à la région centrale de la protéine comprenant notamment le domaine catalytique et différents types de régions variables qui entourent ce cœur structuré, dont les prédictions sont de mauvaise qualité (en orange). Ces régions qui entourent le cœur structuré des protéines ADAMTS-TSL correspondent aux régions N-term et C-term, qui contiennent notamment le propeptide, les domaines/motifs de la région ancillaire et des régions intrinsèquement désordonnées. À noter que les régions intrinsèquement désordonnées sont très présentes au sein des protéines matricielles (dont font partie les ADAMTS-TSL), ce qui leur donne une flexibilité et leur permet alors de se réorganiser pour interagir avec les différents éléments de la matrice extracellulaire [Pey+11].

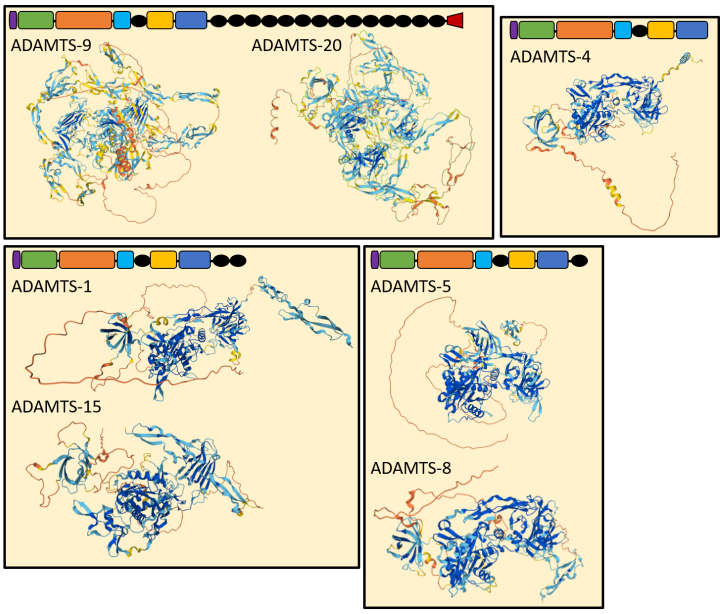

FIGURE 4.4 – **Composition en domaines et structures prédites des hyalectanases humaines**. Les compositions en domaines sont représentées schématiquement comme sur la Figure 4.2. Structures prédites par `AlphaFold` [Jum+21 ; Var+22].

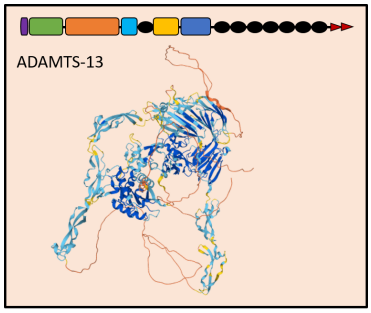

FIGURE 4.5 – **Composition en domaines et structure prédite de ADAMTS-13 humaine.**
La composition en domaines est représentée schématiquement comme sur la Figure 4.2. Structure prédite par `AlphaFold` [Jum+21 ; Var+22].

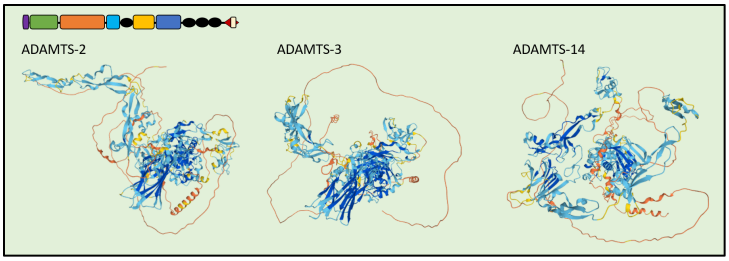

FIGURE 4.6 – **Composition en domaines et structures prédites des procollagènases humaines.**
Les compositions en domaines sont représentées schématiquement comme sur la Figure 4.2. Structures prédites par `AlphaFold` [Jum+21 ; Var+22].

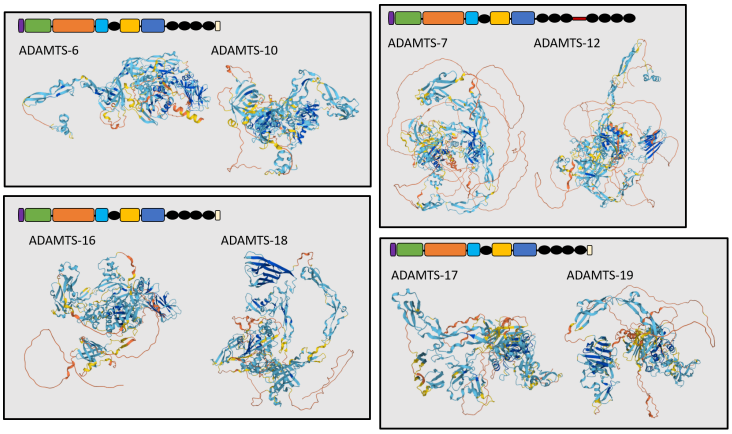

FIGURE 4.7 – **Composition en domaines et structures prédites des ADAMTS humaines associées au réseau de fibrilline-fibronectine.**
Les compositions en domaines sont représentées schématiquement comme sur la Figure 4.2. Structures prédites par `AlphaFold` [Jum+21 ; Var+22].

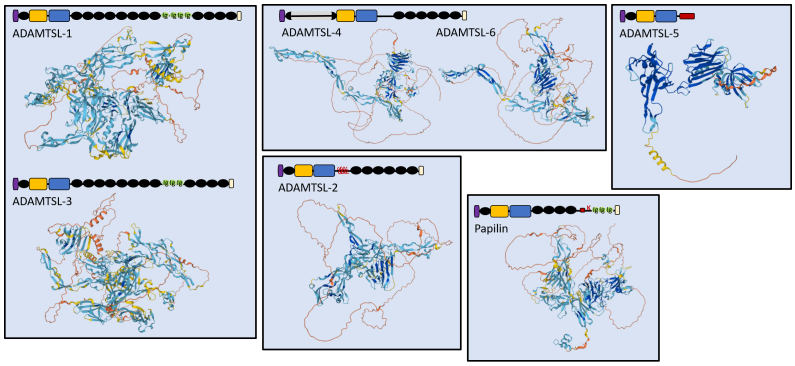

FIGURE 4.8 – **Composition en domaines et structures prédites des ADAMTS-like humaines**. Les compositions en domaines sont représentées schématiquement comme sur la Figure 4.2. Structures prédites par `AlphaFold` [Jum+21 ; Var+22].

4.3 Des protéines multifonctionnelles

Les ADAMTS-TSL sont des composants de la matrice extracellulaire qui constitue le microenvironnement cellulaire. Comme le montre la Figure 4.9 cette matrice forme un réseau moléculaire très complexe qui interagit avec les cellules et régule la communication cellulaire. Les ADAMTS-TSL contribuent à l'homéostasie de la matrice extracellulaire, participent à la stabilité des microfibrilles et régulent de nombreuses fonctions cellulaires comme l'adhésion et la prolifération [Gla+05 ; Mit+05 ; LC11 ; MA18 ; Wan+19b ; Pér+20 ; BW21b].

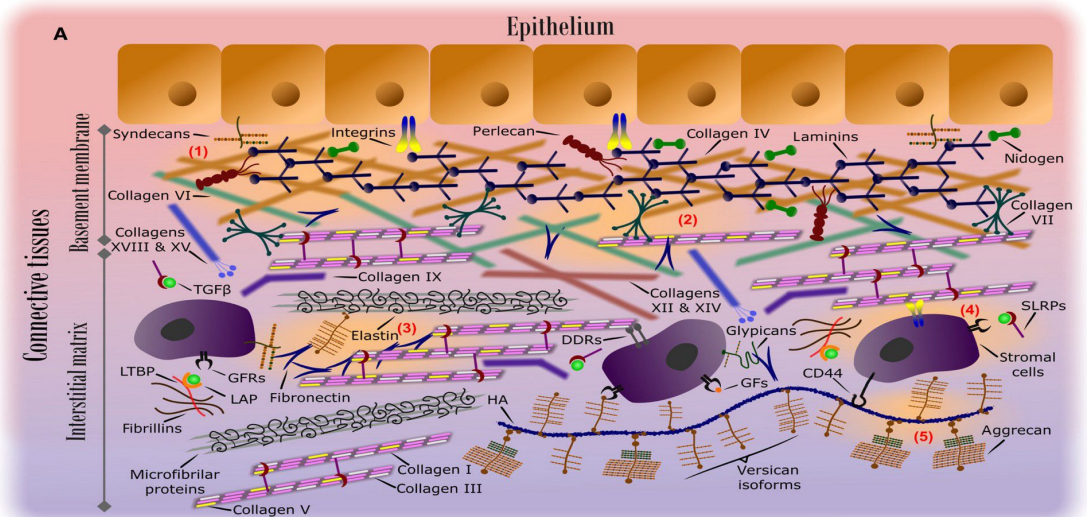

FIGURE 4.9 – **Représentation schématique de l'organisation de la matrice extracellulaire dans l'épithélium et les tissus connectés**, d'après [TMK19]

4.3.1 Des fonctions très variées

L'attribution de « fonction » à une molécule résulte à la fois d'approches globales comme l'association d'expression avec un phénotype, et d'approches ciblées/causales comme la délétion du gène dans un modèle original ou la caractérisation biochimique *in vitro*. Pour identifier les fonctions des ADAMTS-TSL celles-ci ont fait l'objet de très nombreuses études. Elles partagent des effets sur le microenvironnement affectant par exemple la réponse immune (Tableau 4.2) et les troubles musculosquelettiques (Tableau 4.3). De nombreuses revues illustrent cette diversité fonctionnelle des protéines ADAMTS, comme dans le cas des fonctions cardiovasculaires [SG20] et des cancers [CL15].

TABLE 4.2 – **Implication des ADAMTS dans les pathologies/conditions avec une composante inflammatoire**, d'après [Red+21].
n.d. : not determined

ADAMTS	Pathologie/condition	Pro-Anti-inflammatoire	Modèle souris	Type cellulaire	Substrat
ADAMTS1	Aortic aneurysm and dissections	Pro/Anti	Inducible whole-body knockout, Constitutive whole-body heterozygous		Versican
	Atherosclerosis	Pro/Anti		THP-1, macrophages	Versican
	LPS-induced inflammation	Pro			
	Regulation of immune populations	n.d.	Constitutive whole-body knockout		Versican
ADAMTS2	Dermatitis and skin dysfunction	Anti	Constitutive whole-body knockout	Monocytes, macrophages	
ADAMTS4	Aortic aneurysm and dissections	Pro	Constitutive whole-body knockout		Versican
	Atherosclerosis	Pro	Constitutive whole-body knockout	THP-1, macrophages	Versican
	Ischemic stroke and CNS disorders	Anti			
	Osteoarthritis	Pro			Aggrecan
ADAMTS5	Atherosclerosis	n.d.		THP-1, macrophages	
	Influenza virus infection	Pro	Constitutive whole-body knockout		Versican
	Osteoarthritis	Pro			Aggrecan
ADAMTS7	Atherosclerosis	Pro	Constitutive whole-body knockout	THP-1, macrophages	
ADAMTS8	Atherosclerosis	Pro		THP-1	
ADAMTS9	Atherosclerosis	Pro		THP-1	
ADAMTS12	Asthma or allergy	Anti	Constitutive whole-body knockout		
	Inflammatory response	Anti	Constitutive whole-body knockout		
	Osteoarthritis and rheumatoid arthritis	Pro			COMP
ADAMTS13	Autoimmune encephalomyelitis	Anti			
	Multiple sclerosis	Anti			
	Thrombotic thrombocytopenic purpura	Anti			Von Willebrand Factor
	Traumatic microvascular injury	n.d.	Constitutive whole-body knockout		Von Willebrand Factor
ADAMTS14	Skin dysfunction	Anti	Constitutive whole-body knockout		

TABLE 4.3 – **Implication musculosquelettique des troubles associés aux protéases ADAMTS et aux substrats ADAMTS**, d'après [SH21a].

ADAMTS protéase	Implication dans les pathologies humaines	Phénotype musculosquelettique chez la souris *knockout*	Substrats
ADAMTS1	implicated in intervertebral disc (IVD) degeneration	growth retardation without apparent musculoskeletal phenotypes	aggrecan, versican, syndecan 4, TFPI-2, semaphorin 3C, nidogen-1, -2, desmocollin-3, dystroglycan, mac-2, gelatin, amphiregulin, TGF-α, heparin-binding EGF
ADAMTS2	dermatosparaxis Ehlers-Danlos Syndrome (extreme skin fragility, short stature, short hand an feet, joint laxity, craniofacial abnormalities)	dermatosparaxis Ehlers-Danlos syndrome	procollagen type I, II, III, V, Dickkopf-related protein 3, fibronectin, TGF-β RIII, reelin
ADAMTS3	Hennekam lymphangiectasia-lymphedema syndrome	aberrant lymphatic vessel development	procollagen type II, biglycan, LTBP1, fibronectin, TGF-β RIII, reelin, VEGF-C
ADAMTS4	Osteoarthritis (OA) onset, correlated with arthritic severity and progression	not protected from OA	aggrecan, versican, reelin, biglycan, brevican, matrilin-3, α2-macroglobulin, COMP, decorin
ADAMTS5	OA onset, correlated with arthritic severity and progression, implicated in IVD degeneration, suggested role in myoblast fusion delayed onset of arthritis, decreased arthritic severity and progression		aggrecan, versican, reelin, biglycan, matrilin-4, brevican, α2-macroglobulin, decorin
ADAMTS6			LTBP1, syndecan 4
ADAMTS7	tendon maintenance, implicated in rheumatoid and OA heterotopic ossification in tendons, meniscus, and ligaments		COMP, LTBP3, LTBP4
ADAMTS8	implicated in OA		aggrecan
ADAMTS9	nephronophthisis-related ciliopathy (short stature, renal disease) insulin sensitivity in skeletal muscle	soft tissue syndactyly	aggrecan, versican, fibronectin
ADAMTS10	Weill-Marchesani syndrome 1 (short stature, brachydactyly, joint stiffness, hypermuscularity)	shorter long bones, growth plate abnormalities, increased skeletal muscle mass	fibrillin-1, fibrillin-2
ADAMTS12	tendon maintenance, implicated in rheumatoid arthritis and OA	heterotopic ossification in tendons, meniscus, and ligaments	COMP, neurocan
ADAMTS13	thrombotic thrombocytopenic purpura, Upshaw-Schulman syndrome		von Willebrand factor
ADAMTS14			procollagen type I, DKK3, fibronectin, TGF-β RIII
ADAMTS15	implicated in myoblast fusion		aggrecan, versican
ADAMTS16			fibronectin
ADAMTS17	Weill-Marchesani syndrome 4 (short stature, brachydactyly, hypermuscularity)	shorter long bones, growth plate abnormalities, brachydactyly	ADAMTS17
ADAMTS18	microcornea, myopic chorioretinal atrophy and telecanthus, variation in bone mineral density	transient growth delay	
ADAMTS19	non-syndromic heart valve disease	heart valve malformation	
ADAMTS20		soft tissue syndactyly	versican

4.3.2 Association composition domaine/fonction

L'acronyme ADAMTS repose sur la notion même de domaine, *A Disintegrine And Metalloproteisase with ThromboSpondin*. Le domaine metalloprotease qui définit l'appartenance à la famille des protéases à ion métallique, est un domaine dont la fonction, i.e. une activité protéasique, est partagée par toutes les ADAMTS. Près de 150 substrats ont été décrits pour les ADAMTS, un très grand nombre étant partagé par plusieurs ADAMTS (Figure 4.10). C'est le cas notamment pour le groupe des hyalectanases qui ont pour substrat l'aggrecan et le versican. Cependant, les auteurs ont démontré que l'affinité de ces ADAMTS pour leur substrat variait énormément. Ainsi, ADAMTS5 clive très efficacement ces substrats *in vivo* et *in vitro* et est aujourd'hui considéré comme l'aggrécanase physiologique [Sta+05 ; Gla+05 ; Gen+07]. À l'opposé, ADAMTS8 qui possède une organisation en domaine parfaitement identique à celle de ADAMTS5 et qui a une affinité très faible pour ce substrat [Col+04], est aujourd'hui considérée comme une hyalectanase ayant d'autres activités enzymatiques [San+21].

La fonction de ADAMTS5 dans les maladies arthritiques a été directement reliée à son domaine metalloproteinase et sa capacité à dégrader les aggrecans. Par contre, ce domaine metalloprotease chez ADAMTS8 lui procure d'autres fonctions, comme dans l'hypertension artérielle pulmonaire, en clivant l'osteopontin [San+21]. Les auteurs ont proposé que les autres domaines, notamment le « spacer » et les motifs TSP puissent contribuer à la fixation de substrats spécifiques, cependant rien dans l'organisation en domaine de ADAMTS5 et ADAMTS8 ne permet d'identifier ces différences d'interaction. Par ailleurs, la présence d'un domaine metalloproteinase n'est pas suffisante pour prédire la fonction d'une ADAMTS. À titre d'exemple, la protéine ADAMTS12 possède des fonctions indépendantes de son activité catalytique comme son rôle dans l'adhésion et l'invasion des trophoblastes qui impliqueraient des interactions avec les intégrines [BZL11]. Curieusement, alors que le domaine « disintegrin-like » des protéines de la famille ADAM a été largement impliqué dans l'interaction avec les intégrines [BB05], aucune interaction entre le domaine Disintegrin des ADAMTS et les intégrines n'a été identifié. Par contre, le domaine disintegrin-like a été décrit comme nécessaire à la fonction de la protéine ADAMTS13, à savoir le clivage du facteur de Von Willebrand, un substrat unique à cette protease [Gro+09]. De la même façon, le domaine « cystein-rich » présent dans toutes les ADAMTS est indispensable pour cette fonction protéasique unique de ADAMTS13 [GLC15]. Toutes ces observations soulignent à nouveau la limite prédictive des domaines quant à l'identification des fonctions des protéines, et justifient pleinement la recherche

de signatures fonctionnelles plus fines.

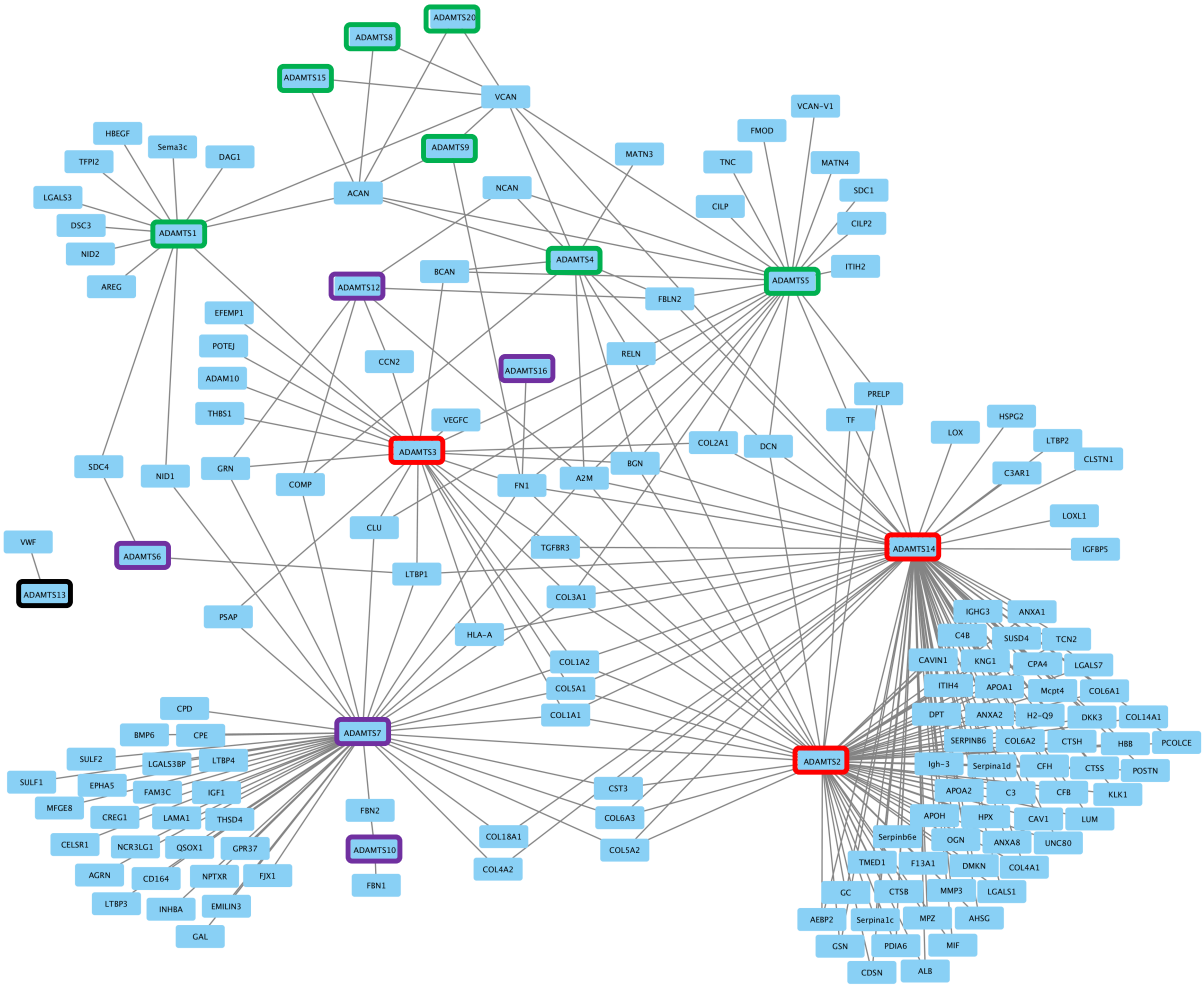

FIGURE 4.10 – **Réseau des ADAMTS et de leurs substrats**.
En vert : le groupe des aggrecanases, en rouge : les procollagenases, en violet : les ADAMTS associées au réseau de fibrillin et fibronectin.

Conclusion

La famille ADAMTS-TSL contient 26 gènes paralogues chez l'humain, issus de forte duplications de gènes au cours de l'histoire des vertébrés. Elle contient les 19 gènes ADAMTS et 7 gènes ADAMTSL (contenant les ADAMTSL ainsi que la Papilin). Les protéines ADAMTS se caractérisent par leur activité et domaine protéolytique, alors que les ADAMTSL ne partagent avec les ADAMTS que la région ancillaire non catalytique. L'étude des ligands et de l'évolution des ADAMTS a permis de les classer en quatre sous-groupes fonctionnels : les procollagénases (ADAMTS-2, -3, -4), la protéase clivant le facteur de Von Willebrand (ADAMTS-13), les hyalectanases (ADAMTS-1, -4, -5, -8, -9, -15, -20) et les couples d'ADAMTS associées aux réseaux de fibrilline/fibronectine (ADAMTS-6/-10, ADAMTS-7/-12, ADAMTS-16/-18, ADAMTS-17/-19). Cependant, cette classification ne rend pas compte de la diversité des fonctions observées au sein d'un même groupe, comme les hyalectanases et les ADAMTS associées aux réseaux de fibrilline/fibronectine.

Deuxième partie

Contributions

L'association domaine/fonction n'étant pas suffisante pour caractériser les spécificités fonctionnelles des 26 protéines ADAMTS-TSL humaines, nous proposons dans cette thèse une approche basée sur la reconstruction de l'histoire évolutive de cette famille de protéines homologues, afin de considérer la modularité de ces protéines multidomaines dû au brassage exonique et l'évolution des différentes fonctions qu'elles partagent. L'utilisation d'alignement partiel local multiple (PLMA) a servi à définir des régions fortement conservées par tous les sous-groupes de séquences existants et ainsi de considérer l'hétérogénéité des protéines/sous-groupes de la famille ADAMTS-TSL. L'utilisation d'une réconciliation mDGS de ces régions conservées a permis de reconstruire leurs évolutions, indépendantes et interdépendantes des gènes et des espèces, et ainsi de les associer à des sous-groupes possédant un même ancêtre commun. Grâce à la reconstruction des scénarios ancestraux des fonctions, nous proposons une nouvelle approche de caractérisation de motifs fonctionnels adaptée aux protéines multidomaines où régions conservées et fonctions sont associées sur la base de leur évolution partagée et leurs acquisitions synchrones au cours de l'histoire évolutive des gènes (Figure 4.11).

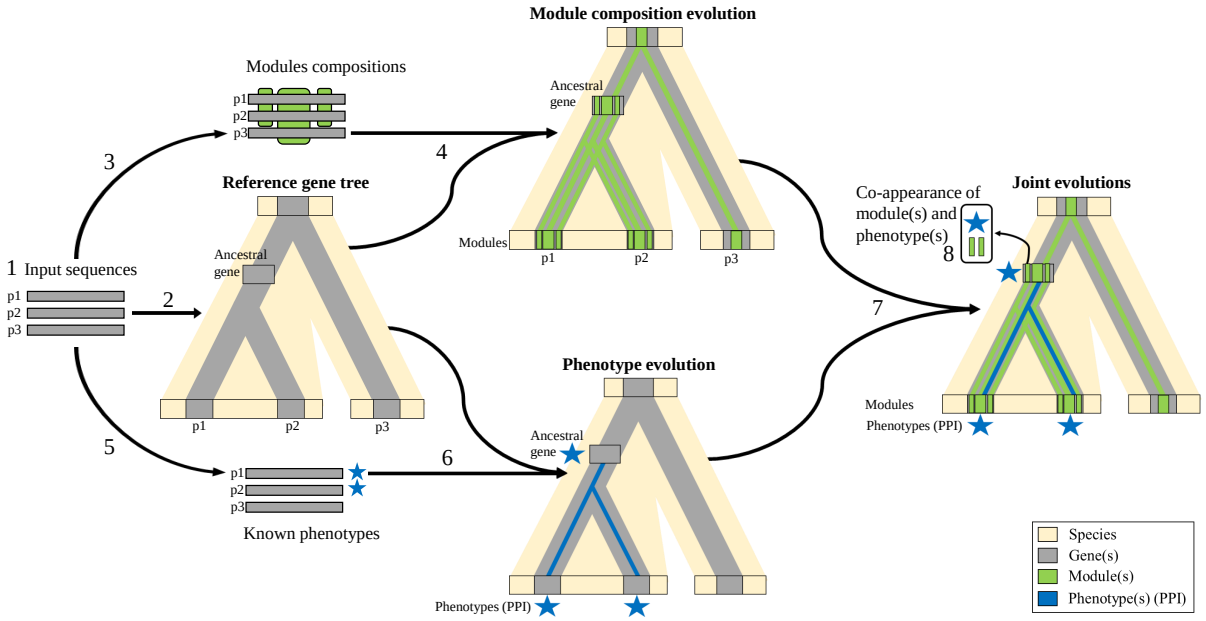

FIGURE 4.11 – **Prédiction phylogénétique de modules fonctionnels.**
Synthèse de la totalité de notre méthode, illustrée ici pour un jeu de séquences contenant deux paralogues p1 et p2 (d'une même espèce) et leur orthologue p3 (d'une autre espèce). Les huit étapes sont les suivantes : 1) Sélection d'espèces et récupération des séquences ; 2) Inférence de l'arbre de référence des gènes à partir des séquences des protéines ; 3) Identification de modules de séquences conservés ; 4) Inférence de l'évolution des compositions en modules des gènes ancestraux de l'arbre de référence des gènes ; 5) Annotation des protéines avec des phénotypes d'intérêts (i.e., les interactions protéines-protéines) ; 6) Reconstruction des scénarios ancestraux des phénotypes à travers l'arbre de référence des gènes ; 7) Jonction des informations d'évolution des modules et des phénotypes, afin d'associer chaque gène ancestral de l'arbre de référence des gènes avec une composition en modules, et une liste de phénotypes ; 8) Prédiction de signatures fonctionnelles par identification d'événements de coapparition de module(s) et de phénotype(s).

Chapitre 5

Phylogénie des ADAMTS-TSL

Les protéines de la famille ADAMTS-TSL ont déjà fait l'objet études phylogénétiques (voir Section 4.1). Cependant, aucune de ces études ne propose d'arbre phylogénétique regroupant ces deux familles, pourtant considérées comme appartenant à une unique superfamille : la superfamille ADAMTS-TSL [Apt20 ; Thé+21]. Nous posons ici l'hypothèse que les ADAMTS et les ADAMTSL possèdent un ancêtre commun, et nous proposons une première phylogénie les regroupant. Cette phylogénie ADAMTS-TSL a pour but de représenter leur évolution, de manière à servir ensuite de modèle sur lequel nous étudierons l'évolution des séquences et des phénotypes de la superfamille, ceci dans le but de caractériser des motifs fonctionnels chez les 26 paralogues ADAMTS-TSL humains. La problématique est alors d'inférer une phylogénie ADAMTS-TSL qui représentera l'évolution des gènes de la lignée humaine, de l'ancêtre *Bilateria* jusqu'à l'humain, et incluant 8 autres espèces animales.

Dans ce chapitre, nous allons présenter nos contributions liées à la construction de jeux de données de séquences de protéines et à l'inférence d'une phylogénie ADAMTS-TSL à partir de ces séquences (Figure 5.1). Dans un premier temps, nous allons étudier les données de séquences disponibles (Section 5.1), et proposer un protocole de récupération des séquences homologues (i.e., orthologues et paralogues) ainsi qu'une méthode pour filtrer les différentes isoformes (Section 5.2). Dans un second temps (Section 5.3), nous proposerons une phylogénie regroupant les ADAMTS et les ADAMTSL au sein d'un seul arbre de référence (Section 5.3.1), ainsi que différentes analyses visant à tester la robustesse de ses différentes bipartitions (Sections 5.3.2 et 5.3.3).

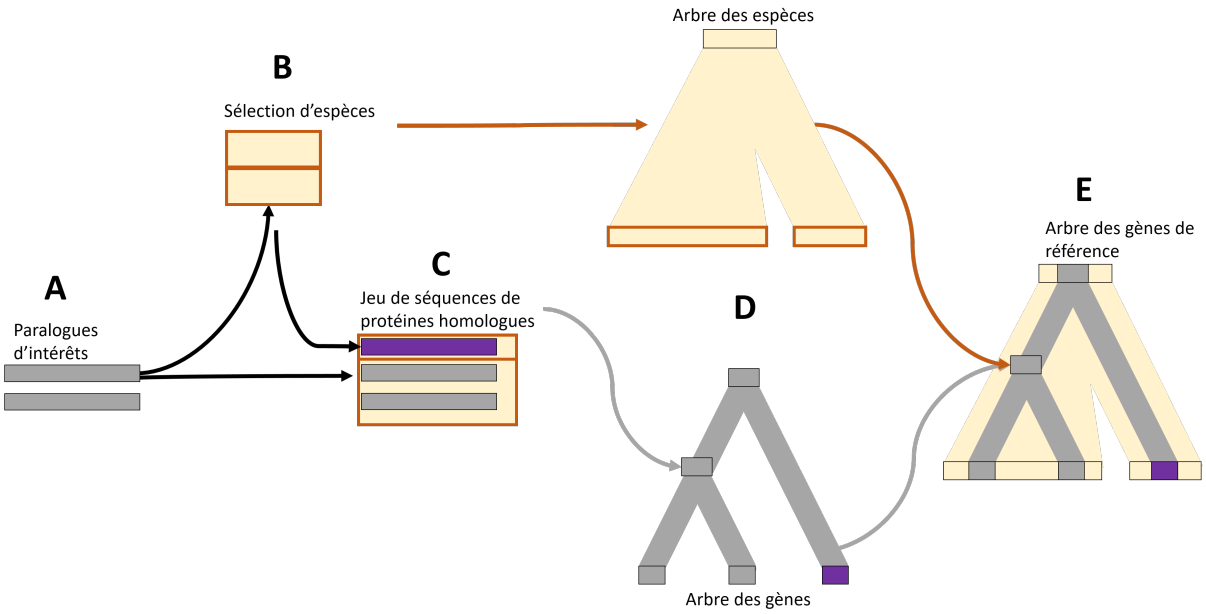

FIGURE 5.1 – **Inférer l'évolution des gènes.**
En partant d'un jeu de séquences paralogues (d'une même espèce) **(A)**, notre but est de sélectionner des espèces dans lesquelles des homologues seraient présents **(B)**, de récupérer ces homologues **(C)**, d'obtenir l'arbre des espèces et l'arbre des gènes **(D)**, de les réconcilier (en tenant compte des spéciations et des duplications paralogues) pour obtenir un arbre des gènes de référence **(E)**.

5.1 Études préliminaires des séquences ADAMTS-TSL

Il n'existe pas de jeu de séquences de référence ADAMTS-TSL suffisamment exhaustif disponible pour notre inférence phylogénétique. C'est pourquoi nous avons choisi d'en constituer un, et donc d'étudier différentes méthodes de récupération de séquences homologues aux ADAMTS-TSL humaines. Nous cherchons à construire un jeu de séquences de référence pour les ADAMTS-TSL.

Cette section regroupe différentes études que nous avons réalisé sur les ADAMTS-TSL, notamment vis-à-vis du contenu de leurs séquences, leur représentation dans le vivant et la disponibilité de leurs séquences. Le but commun à toutes ces études préliminaires était d'obtenir une vision globale des ADAMTS-TSL de manière à élaborer le jeu de données le plus adéquat pour notre étude. Il s'agit ici d'identifier un protocole adapté à la récupération des séquences homologues aux 26 ADAMTS-TSL humaines, chez différentes espèces de métazoaires.

Le jeu de données final devra répondre aux trois critères suivants :

1. **Représenter le contenu en séquences/motifs le plus complet possible**
   Ce qui revient à chercher la séquence la plus complète possible pour un gène (e.g., problème des isoformes et des exons excisés),
2. **Représenter la diversité fonctionnelle des ADAMTS-TSL**
   Ce qui revient à rechercher exhaustivement les homologues de la famille ADAMTS-TSL présents chez un certain nombre d'espèces d'intérêts,
3. **Représenter l'évolution des ADAMTS-TSL humaines**
   Ce qui revient à rechercher des espèces représentatives de l'évolution des ADAMTS-TSL humaines, c'est-à-dire des espèces de plus en plus éloignées de l'homme, d'autres groupes taxonomiques permettant des considérer des ancêtres communs d'intérêts.

5.1.1 Séquences de nucléotides ou d'acides aminés ?

Afin d'identifier la meilleure source de données, nous avons commencé par étudier la cohérence des données concernant les séquences de protéines, des transcrits et des gènes.

Sans aucun *a priori* sur le rôle biologique du type de séquence, nous avons comparé le contenu en résidus des séquences codantes des gènes, des transcrits alternatifs et des protéines, correspondant aux 26 gènes ADAMTS-TSL humains. Les séquences codantes des gènes ont été récupérées dans la base de données *The Consensus Coding Sequences : CCDS* (pour plus de détails, voir 1.1.3.2). Les séquences des transcrits sont issues de la base de données Ensembl. Pour chaque gène, nous avons sélectionné le transcrit contenant le plus grand nombre d'exons. Pour les séquences des protéines, nous avons utilisé les séquences dîtes *canoniques* issues de la base de données Uniprot. Nous avons ainsi trois types de séquences de références, pour chacun des 26 gènes humains ADAMTS-TSL : 1) la séquence codante du gène (CCDS), 2) la séquence du transcrit alternatif contenant le plus d'exons (Ensembl), 3) la séquence de la protéine *canonique* Uniprot (Uniprot). Nous avons ensuite aligné deux à deux ces séquences (codant/transcrit, codant/protéine, transcrit/protéine), dans le but de comparer leur contenu et de mettre en évidence quels résidus/exons sont présents dans quelles séquences. Les séquences nucléotidiques (i.e., codantes, transcrits) ont été traduites en séquences protéiques en utilisant l'outil `Transeq` [RLB00]. Tous les alignements ont ensuite été réalisés avec l'outil `Needle` [NW70]. Les résultats des comparaisons pour les 26 gènes humains sont synthétisés en Figure 5.2.

Pour 20 des 26 gènes, le contenu des trois types de séquences est strictement identique, indiquant une parfaite cohérence entre les exons annotés des gènes, les transcrits alternatifs ayant le plus d'exons et la protéine canonique Uniprot. Pour les six gènes faisant exception, nous avons identifié quatre cas de figure :

1. **Les transcrits et les séquences codantes possèdent une séquence plus longue que celle de la protéine canonique Uniprot**
   Pour les gènes ADAMTSL-4, ADAMTS-14, ADAMTS-19, les séquences codantes des gènes sont cohérentes avec les transcrits, mais sont plus longues que les séquences des protéines canoniques Uniprot (respectivement 23, 3 et 6 acides aminés de plus). Cependant, si on regarde les autres protéines isoformes présentes dans la base de données Uniprot, on retrouve, pour chacun des trois gènes, une autre protéine isoforme qui contient strictement le même contenu en résidus que son transcrit alternatif le plus long et sa séquence codante. Cette observation signifie que la protéine isoforme considérée comme séquence canonique Uniport n'est pas

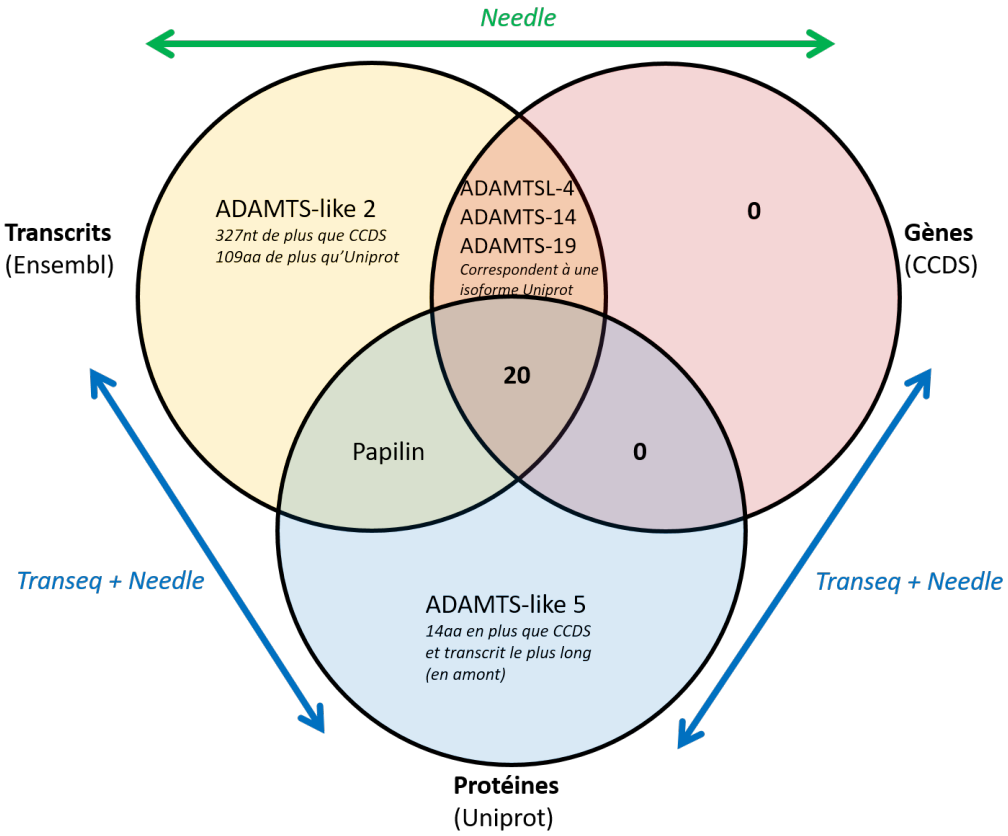

FIGURE 5.2 – **Comparaison des longueurs des séquences codantes CCDS, transcrits Ensembl et protéine canonique Uniprot des 26 ADAMTS-TSL humaines.**
Pour chacun des 26 gènes ADAMTS-TSL humain, la séquence traduite de son gène (région codante), la séquence traduite de son transcrit le plus long et la séquence de sa protéine canonique Uniprot sont alignées avec `Needle`. Les différences de longueurs entre les séquences alignées sont indiquées en nombre de nucléotides (nt) et d'acides aminés (aa).

la séquence de l'isoforme la plus longue, et que d'autres protéines isoformes plus longues sont connues pour ces gènes (cohérents avec les transcrits et séquences codantes).

2. **Le transcrit et la protéine possèdent une séquence plus longue que la séquence génomique**
   Dans le cas de la Papilin, le transcrit et la protéine possèdent 27 acides aminés de plus que la séquence codante proposée par CCDS, suggérant une définition non complète de la séquence codante du gène dans CCDS, où certains exons ne sont pas annotés.

3. **Le transcrit a plus de résidus que la protéine canonique et la séquence codante**
   Pour le cas ADAMTS-like 2, la séquence traduite de son transcrit possédant tous ses exons contient 109 acides aminés de plus que la séquence codante traduite et que la séquence de la protéine. Ces 109 acides aminés correspondent à un exon de 327 nucléotides, dont la présence dans une protéine isoforme n'a pas été observée expérimentalement.

4. **La protéine possède une séquence plus longue que le transcrit et la séquence codante**
   La séquence de la protéine canonique Uniprot d'ADAMTS-like 5 possède 14 acides aminés de plus que celle de son transcrit traduit ou de sa séquence codante traduite. Ces 14 acides aminés supplémentaires correspondent à une initiation de traduction en amont de celle annotée dans les données CCDS.

À noter que, pour aucun des 26 gènes, la séquence codante annotée dans CCDS ne possède plus de résidus que les deux autres. De plus, la moitié des différences provient du choix de l'isoforme canonique d'Uniprot, qui n'est pas nécessairement le plus long et le plus représentatif du contenu en exons d'un gène. L'autre moitié des différences est issue du type de donnée, que ce soit la possibilité d'un transcrit alternatif jamais observé ou d'une protéine isoforme observée dont l'initiation n'est pas annotée comme telle sur les données génomiques CCDS. L'objectif de nos travaux est d'identifier des motifs fonctionnels au sein des protéines ADAMTS-TSL, nous avons donc utilisé les séquences protéiques plutôt que les séquences des transcrits alternatifs. Cependant, l'utilisation des séquences protéiques nécessite d'analyser les multiples protéines isoformes.

5.1.2 Taxonomie des espèces

Tout au long de cette thèse, et particulièrement dans ce chapitre, nous utilisons des phylogénies des espèces afin de représenter les données liées aux différentes espèces, ou de corriger des arbres de gènes. Nous représentons les espèces par leur *taxid* tels que définis dans la taxonomie du NCBI (e.g., *Homo sapiens* est représenté par *9606*). Nous récupérons ensuite à partir de *NCBI taxonomy* [Sch+20] l'arbre phylogénétique des espèces correspondantes aux *taxid* d'intérêts.

5.1.3 Méthodes préliminaires de construction du jeu de séquences

Afin de récupérer les homologues des ADAMTS-TSL humaines, il est possible d'utiliser les annotations (i.e., protéines référencées comme ADAMTS-TSL), la similarité de séquences (e.g., par recherche via `BLAST RBH`, détaillé en 5.1.3.2), ou des méthodes plus complexes également basées sur la similarité de séquences (e.g., `OrthoFinder`[EK15 ; EK19]). Le but ici est d'identifier les espèces possédant des gènes/protéines homologues (orthologues et paralogues) aux 26 ADAMTS-TSL humaines (i.e., issues d'un gène ancestral commun), et de récupérer ces séquences homologues. Pour ceci, nous nous sommes intéressés à différentes méthodes de récupération de séquences homologues, en étant exhaustif et de manière à retracer leur évolution jusqu'à l'humain.

Nous testerons trois méthodes, toutes basées sur une vision matricielle de l'homologie, qui consiste à associer à un paralogue humain l'homologue d'une autre espèce, le plus proche en termes de similarité de séquences. La première (5.1.3.1) consiste à chercher l'homologue d'un paralogue humain, via ses annotations (e.g., nom, description, composition en domaines). La seconde (5.1.3.2) se passe de ces *a priori* et cherche l'homologue le plus proche en termes de similarité de séquences (par `BLAST RBH`). La troisième (5.1.3.3) vise à améliorer l'exhaustivité de la deuxième approche, en utilisant la similarité de séquences de manière itérative. Motivés par les limites de ces différentes approches, nous opterons pour une méthode basée sur l'outil `OrthoFinder` (5.2).

5.1.3.1 Jeu de séquences issues d'annotations

Une grande partie des protéines disponibles dans la base de données `Uniprot` possèdent des annotations précises, comprenant leur nom, une description, leur composition en domaines connus, ainsi qu'un certain nombre d'autres informations. Le nom (e.g., `Adamts1`, `ATS1 metalloproteinase`, `ADAM metallopeptidase with thrombospondin type 1 motif 1`) d'une protéine donne généralement des informations sur une relation d'orthologie ou d'homologie, nous permettant de récupérer les séquences homologues par la recherche d'annotations « ADAMTS » ou « ADAMTS-like ».

Pour une sélection de 19 espèces, nous avons récupéré les protéines canoniques `Uniprot` possédant une annotation « ADAMTS », « ATS », « ADAMTSL » ou « ATL » dans leur nom. De plus, nous avons complété la matrice obtenue (Figure 5.3) avec les protéines comprenant une composition en domaines similaire aux ADAMTS-TSL humaines. Nous obtenons ainsi 341 séquences de protéines dont les annotations indiquent que ce sont des homologues des ADAMTS-TSL humaines. Les différentes annotations nous permettent d'associer chaque homologue à la protéine humaine qui devrait lui correspondre (rangée par colonne sur la Figure 5.3). Les critères d'associations sont : le partage d'un même nom de protéine, d'un même nom de gène, d'une même composition en domaines. Cependant, toute protéine inconnue ou sans annotation ne sera pas récupérée, ce qui revient à considérer que les protéines non annotées, bien que pouvant être similaires aux séquences humaines, ne sont pas homologues aux ADAMTS-TSL humaines. Bien que potentiellement non exhaustif, ce premier jeu de séquences nous permet d'avoir un premier aperçu de l'expansion paralogue de la famille ADAMTS-TSL. En particulier chez l'ancêtre des *Gnathostomata* (vertébrés à mâchoires), où le nombre de copies (lignes sur la Figure 5.3) par espèce augmente considérablement, l'espèce de tunicié *Ciona intestinalis* possédant 8 ADAMTS-TSL, contre 17 pour *Callorhinchus milii* (le vertébré *Gnathostome* en possédant le moins).

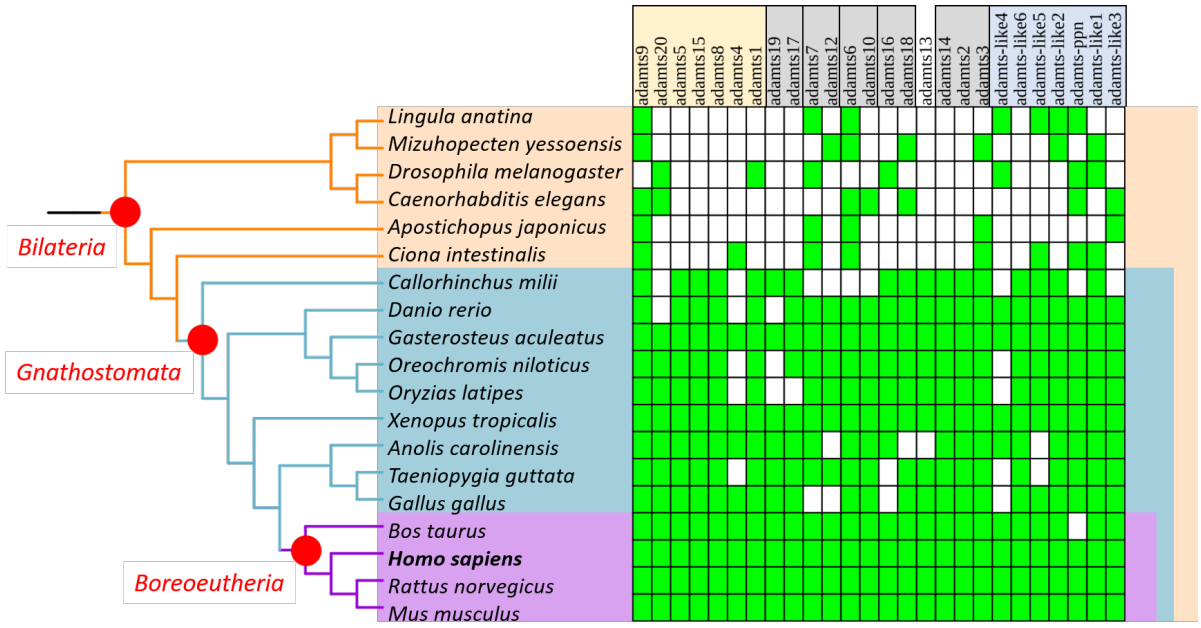

FIGURE 5.3 – **Représentation en matrice des 341 ADAMTS-TSL annotées comme homologues selon les annotations de 19 espèces.**
Chaque séquence ADAMTS-TSL annotée comme telle est représentée par un carré vert, en fonction de l'espèce dont elle provient (ligne), et de la séquence humaine la plus similaire en annotations (colonne).

5.1.3.2 Jeu de séquences issues de relations d'orthologies

Les annotations de séquences sont limitées par la connaissance actuelle et les descriptions établies en amont. Mais il est également possible de rechercher des séquences homologues sans *a priori*, par similarité. La méthode la plus courante consiste à réaliser un alignement de séquences appelé `BLAST RBH` (*Reciprocal Best Hits*), ou `BLAST` réciproque [ML08]. Un `BLAST RBH` consiste en l'alignement d'une séquence avec le protéome d'une espèce dans laquelle on cherche une protéine homologue. Les protéines les plus similaires à la séquence de départ (s'alignant le mieux, et donc possédant les scores les plus hauts, *best hit*) sont ensuite alignées contre le génome dont est issue la protéine de départ. Si la protéine de départ est retrouvée, le *best hit* est considéré comme réciproque, et donc comme une protéine homologue. Dans le cas où le *best hit* n'est pas réciproque, il est considéré comme trouvé par hasard. Nous avons choisi de ne pas réimplémenter de `BLAST` réciproques, ni d'en recalculer, mais d'utiliser des relations d'orthologies déjà calculées sur ce principe et stockées dans la base de données d'orthologie `OrthoInspector 3.0` [Nev+19]. Notre but est ici de rechercher des homologues aux 26 protéines ADAMTS-TSL humaines, via les relations d'homologies présentes dans `OrthoInspector 3.0`.

Sur la base des relations d'orthologies RBH d'`OrthoInspector 3.0`, nous avons étudié la présence d'homologues des 26 protéines ADAMTS-TSL humaines au sein de 48 espèces de métazoaires. Le nombre d'homologues pour chacune de ces espèces est présenté sur la Figure 5.4. Nous en avons tiré deux observations d'intérêt : 1) il existe des homologues ADAMTS-TSL humaines au sein de la quasi-totalité des métazoaires, 2) le nombre estimé de copies paralogues d'ADAMTS-TSL varie grandement au cours de l'évolution des métazoaires, allant de 1 à 26, et augmente particulièrement chez l'ancêtre *Chordata* (en bleu sur l'arbre) et l'ancêtre *Euteleostomi*.

Ces observations s'ancrent dans une logique d'apparition et de développement de la matrice extracellulaire (MEC) chez les métazoaires. En effet, la présence d'une matrice extracellulaire riche en collagène est une caractéristique majeure des métazoaires. La présence d'ADAMTS-TSL, protéines régulatrices de cette matrice extracellulaire au sein de l'ensemble des métazoaires, a ainsi beaucoup de sens. Les ADAMTS-TSL sont présentes dès lors que l'organisme possède une matrice extracellulaire [Bru+15]. De plus, l'estimation du nombre de copies par organisme peut s'expliquer par la complexité de la matrice extracellulaire des différents organismes [Hyn12] ; plus un organisme possède une matrice extracellulaire complexe, plus il possède de copies paralogues d'ADAMTS-TSL. L'apparition et l'évolution de la famille ADAMTS-TSL (en nombre de copies) sont ainsi

étroitement liées à l'apparition et à la complexification de la matrice extracellulaire au sein des métazoaires [Hux+05]. Cette augmentation du nombre de copies résulte entre autre des duplications complètes de génomes chez les vertébrés [Bru+15].

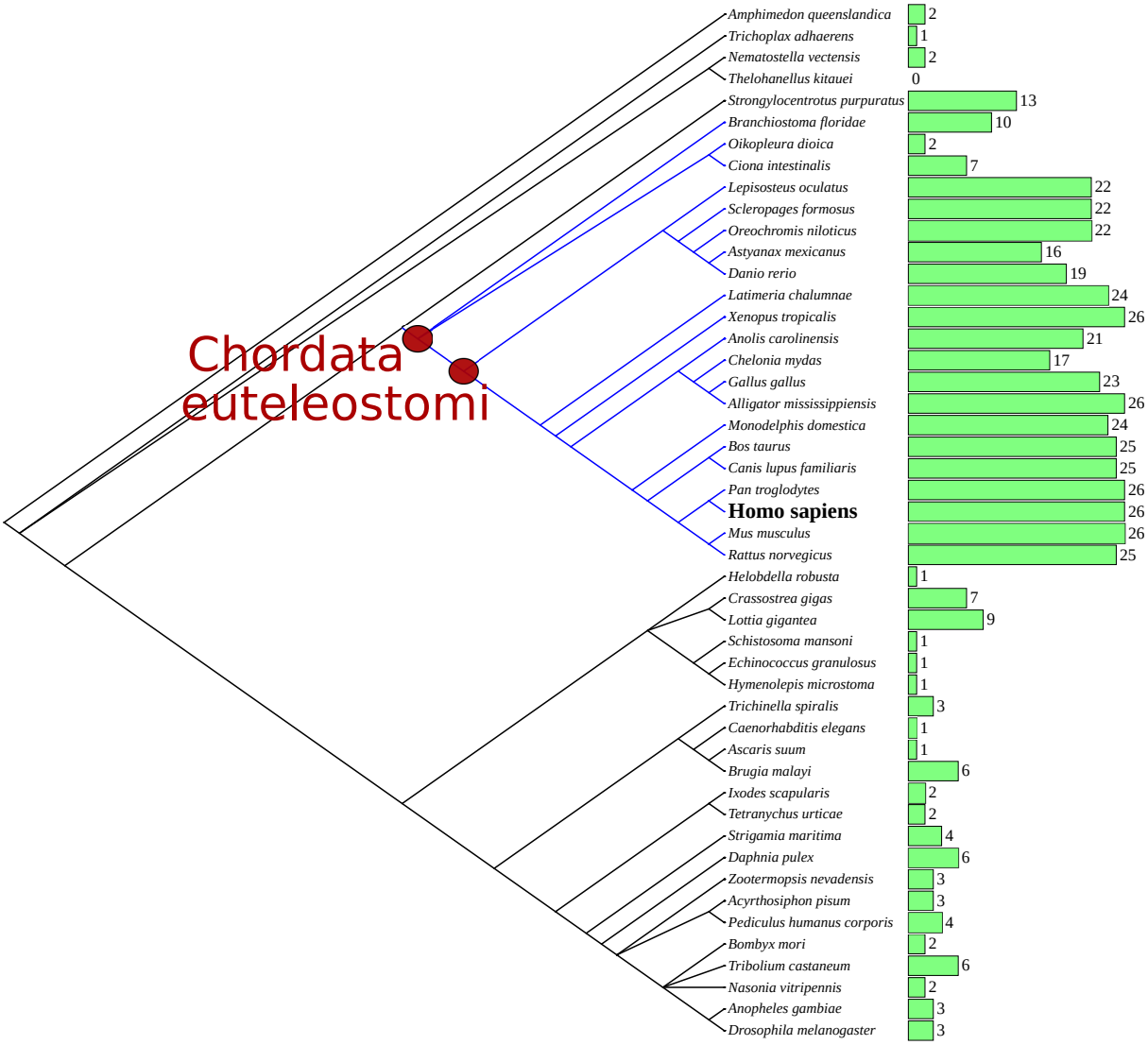

FIGURE 5.4 – **Estimation du nombre d'homologues des ADAMTS-TSL humaines au sein de 48 espèces de métazoaires grâce à Orthoinspector.**
L'arbre phylogénétique des 48 espèces est extrait de la *taxonomie NCBI* [Sch+20]. Le nombre d'ADAMTS-TSL par espèce est estimé à partir des homologues connus des 26 ADAMTS-TSL humaines dans `OrthoInspector 3.0` [Nev+19]. L'arbre est visualisé avec l'outil `Itol` [LB19a].

Nous venons de nous intéresser aux homologues directs des protéines humaines chez une autre espèce B (et ceci pour 48 espèces). Cependant, rien ne nous assure que toutes

les ADAMTS-TSL de cette espèce B sont recensées comme ayant une relation d'orthologie avec les séquences humaines, en particulier si cette espèce B est lointaine de l'espèce humaine. Par exemple, dans `OrthoInspector 3.0`, la papilin d'*H. sapiens* a pour homologue la papilin de *D. melanogaster*, mais pas celle de *C. elegans*. Mais si nous regardons les homologues de la papilin de *D. melanogaster*, cette dernière a pour homologue la papilin de *C. elegans* (Figure 5.5). La relation entre la papilin *H. sapiens* et la papilin *C. elegans* n'est pas présente dans la base de données `OrthoInspector 3.0`, mais peut être déterminée transitivement en considérant les espèces intermédiaires. Cet exemple suggère qu'utiliser les différentes espèces, et effectuer des `BLAST` réciproques entre toutes les espèces intermédiaires (non-humaines) pourrait nous permettre de constituer un jeu de séquences plus exhaustif.

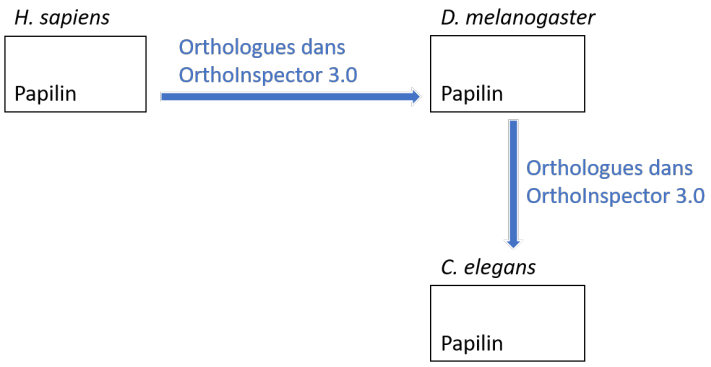

FIGURE 5.5 – **Relation d'orthologie de la papilin dans OrthoInspector 3.0.**
Parmi les relations d'orthologies présentes dans la base de données d'orthologues `OrthoInspector 3.0` [Nev+19], il existe une relation d'orthologie entre la papilin de *H. sapiens* et la papilin de *D. melanogaster*, ainsi qu'une relation d'orthologie entre la papilin de *D. melanogaster* et la papilin de *C. elegans*. Mais il n'y a pas de relation d'orthologie entre la papilin de *H. sapiens* et la papilin de *C. elegans*.

Par l'utilisation de similarités réciproques entre les séquences issues de `BLAST RBH` et disponibles dans la base de données `OrthoInspector 3.0`, nous avons pu construire un jeu de données de 523 séquences homologues ADAMTS-TSL, au sein de 48 espèces. Ce jeu de données nous a permis de conforter nos observations sur l'augmentation du nombre de paralogues au cours de l'évolution de la lignée humaine, mais possède uniquement des séquences homologues aux 26 paralogues humains.

5.1.3.3 Relations indirectes et recherche itérative d'homologues

La recherche d'homologie basée uniquement sur les homologues des protéines humaines selon `OrthoInspector 3.0` au sein d'autres espèces peut omettre des relations d'homologies. Comme nous venons de le voir avec l'exemple de la papilin *H. sapiens* et de *C. elegans*, mais peut être inférée transitivement par leurs relations d'homologies communes avec la papilin de *D. melanogaster* (espèce intermédiaire). C'est pourquoi nous avons choisi ici d'explorer les relations indirectes issues d'espèces intermédiaires (relations d'homologies transitives), afin de constituer un jeu de séquences homologues le plus exhaustif possible.

Nous avons implémenté une recherche itérative d'homologues de la famille ADAMTS-TSL, qui cherchent les homologues des 26 protéines ADAMTS-TSL humaines, avant de chercher les séquences homologues aux séquences homologues trouvées de manière transitive. Décrit en Figure 5.6, le principe de cette méthode est simple, 1) nous recherchons les homologues de nos protéines de départ, avant de 2) rechercher les homologues des homologues nouvellement identifiés (première itération). Cependant, si nous itérons trop de fois, nous nous retrouvons à considérer des protéines très divergentes des protéines de départ. Les relations d'homologies sont toutes issues de la base de données `OrthoInspector 3.0`. Nous proposons ici de les parcourir de manière itérative.

En seulement quatre itérations, chacune des 48 espèces est associée à 26 séquences homologues (une séquence homologue pour chacune des séquences humaines) ce qui ne semble pas réaliste. Ce qui suggère que ces homologues ne soient pas nécessairement des ADAMTS-TSL, mais des protéines beaucoup plus divergentes appartenant à d'autres familles. De plus, nous avons également récupéré de nouvelles espèces où des séquences homologues sont identifiées.

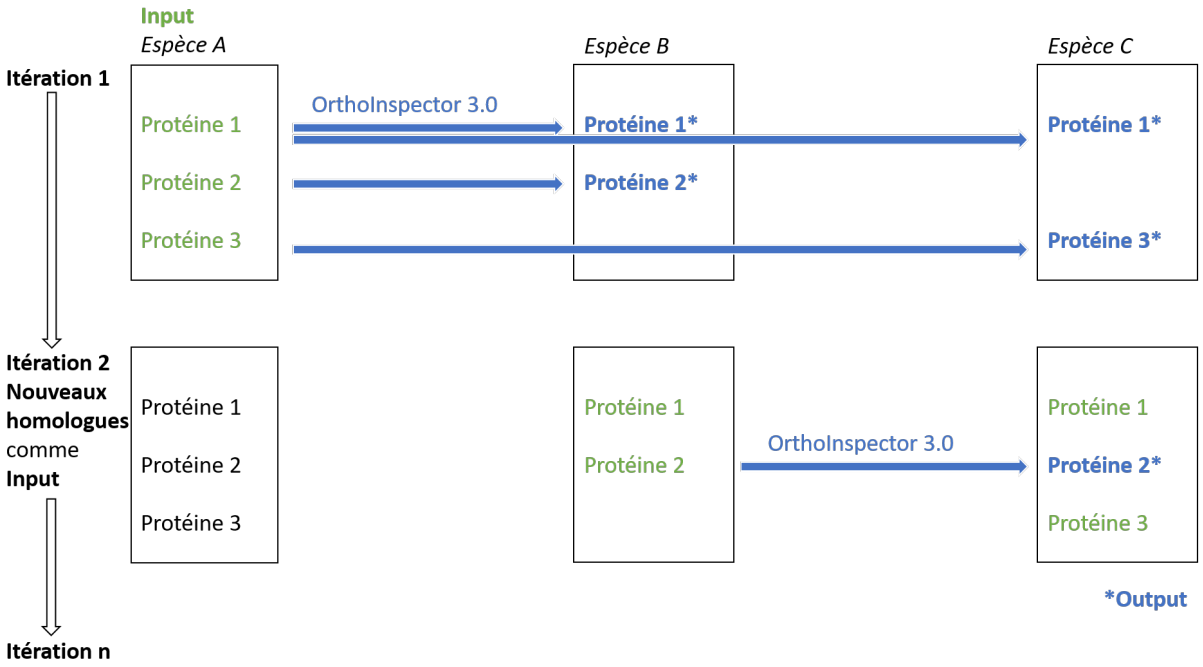

FIGURE 5.6 – **Recherche itérative d'homologues dans la base de données OrthoInspector.** Une première itération à partir des protéines d'intérêts (Itération 1, en vert) permet d'obtenir leurs séquences homologues (Itération 1, en bleu). Une seconde itération permet de chercher les séquences homologues de ces homologues (Itération 2, en bleu), et ainsi d'obtenir de nouveaux homologues (Itération 2, en vert). Itérer permet ainsi de récupérer des séquences homologues de nos protéines d'intérêts plus divergentes.

L'exemple le plus marquant d'identification d'homologues très divergents est celui de deux homologues hors des métazoaires (Figure 5.7). En effet, (itération 1) ADAMTS-1 *Homo sapiens* a pour homologue ADAMTS-1 *Rattus norvegicus*, (itération 2) et ADAMTS-1 *Rattus norvegicus* a pour homologue deux protéines inconnues de *Fonticula alba*[1] (Uniprot : A0A058ZAP4, A0A058ZC40) qui ne fait pas partie des *métazoaires*. Cependant, ces deux protéines de *Fonticula alba* font respectivement 603 et 619 acides aminés, et possèdent toutes les deux uniquement un domaine catalytique M12B. Elles ne possèdent aucun autre des domaines caractéristiques des ADAMTS-TSL (i.e., désintégrine, thrombospondine, spacer). Nous avons ainsi uniquement une correspondance sur les domaines de ces homologues de *F. alba*, mais rien ne prouve qu'elles possèdent un lien de parenté avec les ADAMTS-TSL humaines, et qu'il ne s'agit pas d'un faux positif. Si nous avions considéré *Rattus norvegicus* comme espèce de départ, nous aurions obtenu ces faux positifs en première itération, directement parmi les homologues identifiés par `OrthoInspector 3.0`.

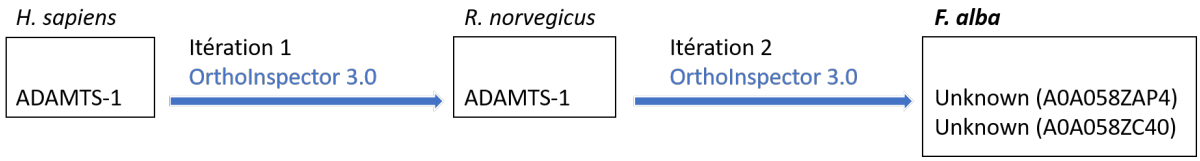

FIGURE 5.7 – **Identification d'homologues d'ADAMTS-TSL chez *Fonticula alba***

Utiliser la similarité de séquences (`BLAST` réciproques) en considérant les relations d'homologies entre toutes les espèces considérées permet d'être plus exhaustif sur les homologues obtenus. Cependant, les relations d'homologies disponibles dans la base `OrthoInspector 3.0` ne permettent pas d'obtenir un résultat satisfaisant. Dans le but de constituer un jeu de séquences de références pour les ADAMTS-TSL, nos études préliminaires identifient la nécessité de :

- Considérer les différentes séquences isoformes
- Utiliser la similarité de séquences pour rechercher les homologues
- Considérer les relations d'homologies entre toutes les espèces
- Utiliser des espèces représentant l'évolution de la lignée humaine

1. *Fonticula alba* est une espèce particulière au sein des *Opisthocontes* (super-règne comprenant notamment les *métazoaires* et les *fongi*). Cette amibe possède un cycle de vie particulier, alternant un comportement unicellulaire (cycle solitaire) et un comportement multicellulaire (cycles sociaux). Elle est étudiée dans le but de comprendre l'origine de l'organisation multicellulaire des *opisthocontes* ainsi que sa divergence avec les *métazoaires* et les *fongi* [Tor+22].

5.2 Construction d'un jeu de séquences considérant orthologues, paralogues et isoformes

Afin de prendre en compte les problèmes décrits dans les parties précédentes, nous avons choisi d'utiliser le logiciel `OrthoFinder` [EK15 ; EK19] pour déterminer les ADAMTS-TSL homologues d'un ensemble d'espèces. Cet outil va nous permettre de diviser les séquences des protéomes d'espèces d'intérêts en groupes de séquences homologues, appelés *Orthogroupes* et qui regroupent les séquences orthologues, paralogues et isoformes de la famille ADAMTS-TSL. Le choix de ce programme soulève deux problèmes à résoudre :

- La sélection d'espèces représentantes de l'évolution des ADAMTS-TSL de la lignée humaine (i.e., de l'ancêtre bilatérien jusqu'à l'homme), et possédant des protéomes et des génomes annotés
- Le traitement des isoformes : représentantes de multiples instances d'un même gène, il s'agit de les regrouper et de sélectionner une isoforme représentative par gène

Cette section présente les différentes étapes de la construction du jeu de séquences ADAMTS-TSL, ainsi que le jeu de séquences homologues obtenu.

5.2.1 Disponibilité des génomes annotés, des protéomes complets et sélection d'espèces

Le programme `OrthoFinder` nécessite en entrée les protéomes des espèces à étudier. Afin de recenser de manière exhaustive les membres de la famille ADAMTS-TSL, nous avons besoin des protéomes les plus complets possibles. Pour regrouper les isoformes par gène, nous aurons également besoin des génomes, si possible annotés. Nous nous intéressons à caractériser les protéines humaines, et pas à déterminer toute l'histoire des ADAMTS-TSL. Nous considérons donc les espèces permettant de caractériser les ancêtres communs de la lignée humaine spécifiquement. Notre objectif est alors de sélectionner des espèces représentantes de l'évolution des ADAMTS-TSL humaines et possédant un protéome et un génome annoté.

Dans le cadre de la sélection d'espèces d'études, nous nous focalisons ici sur la disponibilité des séquences (génome/protéome). L'étude que nous proposons s'intéresse aux assemblages de génomes du *NCBI* [Say+19]. Nous regardons trois éléments ;

1. **La disponibilité d'un génome/protéome**
   i.e., présence d'un assemblage
   Nous estimons la disponibilité d'un assemblage par la présence d'un assemblage référencé par le préfixe GCA dans la base de données *GeneBank* [Say+19].
2. **La qualité d'annotation de l'assemblage**
   Ces assemblages ne sont pas tous annotés. La présence ou non de cet assemblage référencé par le préfixe GCF dans la base de données *RefSeq (NCBI Reference Sequence Database)* indique la présence d'annotations. En effet, si un assemblage GCA possède un niveau d'annotations suffisant, il peut être sélectionné pour devenir un assemblage GCF.
3. **Le niveau de finition de l'assemblage**
   i.e., à quel point les lectures issues d'un séquençage sont assemblées pour former le génome sous forme de chromosomes
   Pour les assemblages RefSeq (GCF), nous considérons également le niveau de finition de l'assemblage, les niveaux d'assemblages étant par ordre croissant de finition : *contig* (regroupements de lectures de séquençage chevauchantes), *scaffold* (regroupement de *contig* par l'utilisation d'informations supplémentaires), *chromosome* (regroupement des lectures de séquençage allant jusqu'à la résolution d'un chromosome), *complet* (toutes les lectures de séquençage sont assemblées de manière à former autant de séquences qu'il existe de chromosomes). Les assemblages *complets* étant rares (ici, seul *C. elegans* possède un assemblage *complet*), nous sélectionnons les assemblages GCF dont le niveau de finition est au moins au niveau des *chromosomes*.

L'analyse de ces trois critères pour nos 48 espèces métazoaires d'intérêt est résumée sur la Figure 5.8. La présence d'un assemblage GCA est indiquée en rouge, la présence d'un assemblage GCF en bleu, et si ce GCF a une finition au niveau chromosomique, sa présence est indiquée en vert. Toutes nos espèces possèdent au minimum un assemblage GCA, et cinq de ces assemblages ne sont pas sélectionnés comme GCF. Parmi ces 40 espèces possédant un assemblage GCF, seules 29 ont un assemblage dont la finition est au niveau des *chromosomes*, les autres possédant au mieux un niveau de finition au niveau des *scaffolds*. Des assemblages satisfont nos critères dans la très grande majorité des clades

de notre arbre des espèces. Les génomes les mieux annotés possédant un meilleur niveau de finition sont principalement des génomes d'espèces appartenant au phylum *Chordata* (sous-arbre bleu).

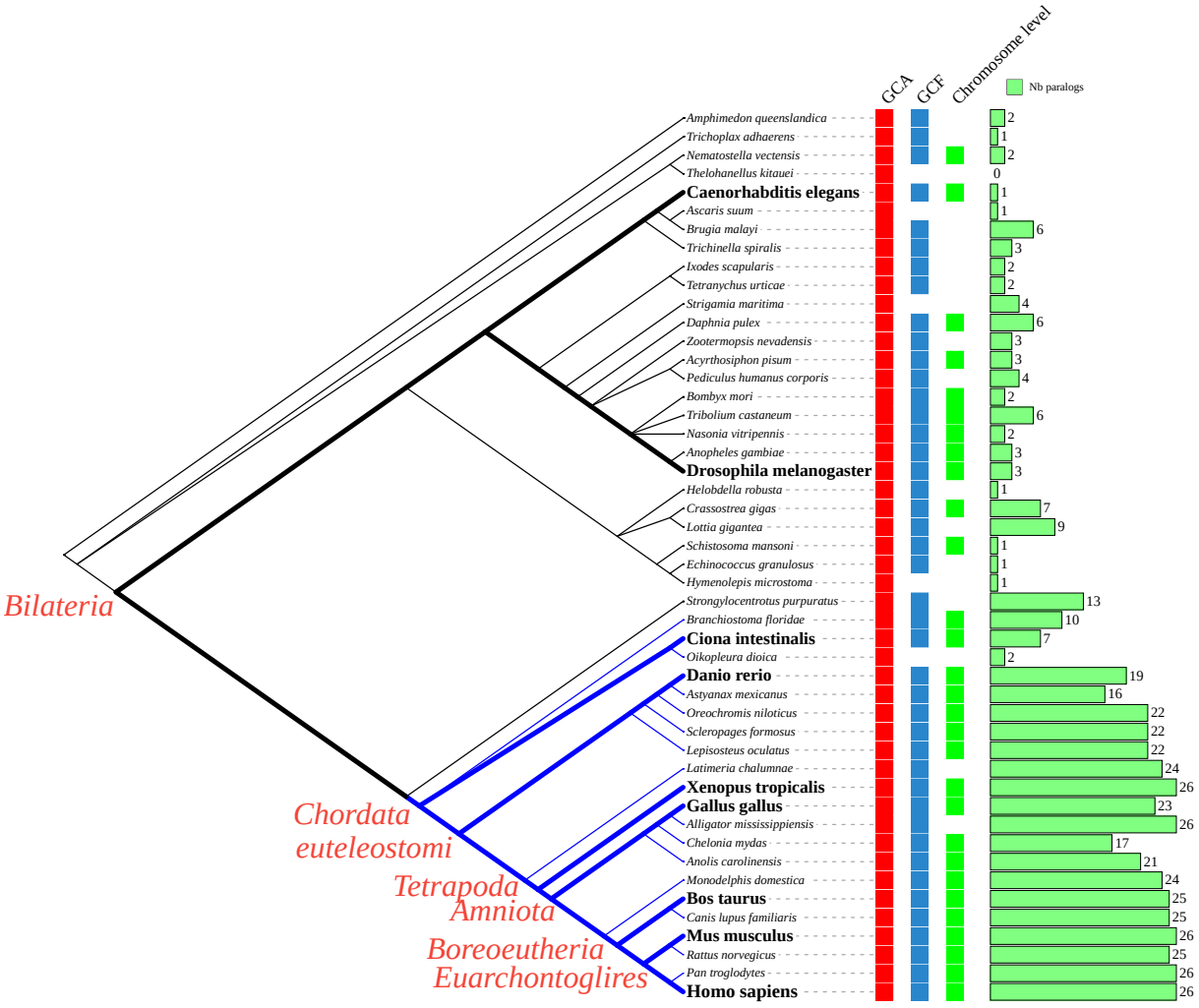

FIGURE 5.8 – **Disponibilité d'assemblages, d'annotations et leurs niveaux de finition, pour 48 espèces de métazoaires**
L'arbre interactif est disponible **sur ce lien**. La disponibilité d'un assemblage GCA est indiquée en rouge, la disponibilité d'un assemblage GCF (et donc d'annotations) est indiquée en bleu. En vert est indiqué la disponibilité d'un assemblage GCF assemblé au niveau chromosomique (ou complet). Le nombre de séquences homologues présenté en Figure 5.4 est également indiqué par les histogrammes verts. Les 9 espèces sélectionnées sont indiquées en gras, et les ancêtres de la lignée humaine sont indiqués aux nœuds internes de l'arbre (e.g., *Chordata* est l'ancêtre du sous arbre coloré en bleu)

Parmi nos 48 espèces représentatives de l'évolution des métazoaires, 29 possèdent une disponibilité de séquences considérée comme suffisante (assemblage GCF annoté, avec finition au niveau des chromosomes) pour la suite de nos analyses. Nous sélectionnons 9 de ces 29 espèces dans le but de pouvoir inférer les ancêtres d'intérêts de l'évolution des ADAMTS-TSL de la lignée humaine (Figure 5.8 et Tableau 5.1).

TABLE 5.1 – **Les 9 espèces, leur taxid et dernier ancêtre commun avec l'homme.**

Espèce	Taxid	Ancêtre avec *H. sapiens*
Homo sapiens	9606	
Mus musculus	10090	*Euarchontoglires*
Bos taurus	9913	*Boreoeutheria*
Gallus gallus	9031	*Amniota*
Xenopus tropicalis	8364	*Tetrapoda*
Danio rerio	7955	*Euteleostomi*
Ciona intestinalis	7719	*Chordata*
Drosophila melanogaster	7227	*Bilateria*
Caenorhabditis elegans	6239	*Bilateria*

Nous avons sélectionné les 9 espèces suivantes (Tableau 5.1) : *Homo sapiens*, *Mus musculus*, *Bos taurus*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Ciona intestinalis*, *Drosophila melanogaster*, *Caenorhabditis elegans*, comme espèces représentantes de l'évolution des ADAMTS-TSL de la lignée humaine et possédant des protéomes et des génomes annotés de qualité.

5.2.2 Récupération d'homologues ADAMTS-TSL avec OrthoFinder

Nous proposons ici un protocole qui permet de constituer un jeu de séquences de protéines regroupant les homologues (orthologues, paralogues) et leurs différentes isoformes, et ceci, sans *a priori*. Ce protocole est basé sur un *clustering* complet des protéomes d'espèces d'intérêt avec le logiciel `OrthoFinder` [EK15 ; EK19]. Le protocole ne nécessite que deux entrées : 1) une liste d'espèces (et leurs protéomes), et 2) une liste de protéines d'intérêt (appartenant aux protéomes des espèces d'intérêt), et permet d'obtenir un jeu

de séquences homologues et de protéines isoformes. Cette méthode peut être résumée en trois étapes représentées en Figure 5.9 : 1) Les protéomes sont divisés en *orthogroupes* avec `OrthoFinder`, 2) sélection des *orthogroupes* contenant au moins l'une des protéines d'intérêt, 3) Le jeu de données de séquences est construit en agrégeant les séquences des protéines présentes dans les *orthogroupes* sélectionnés.

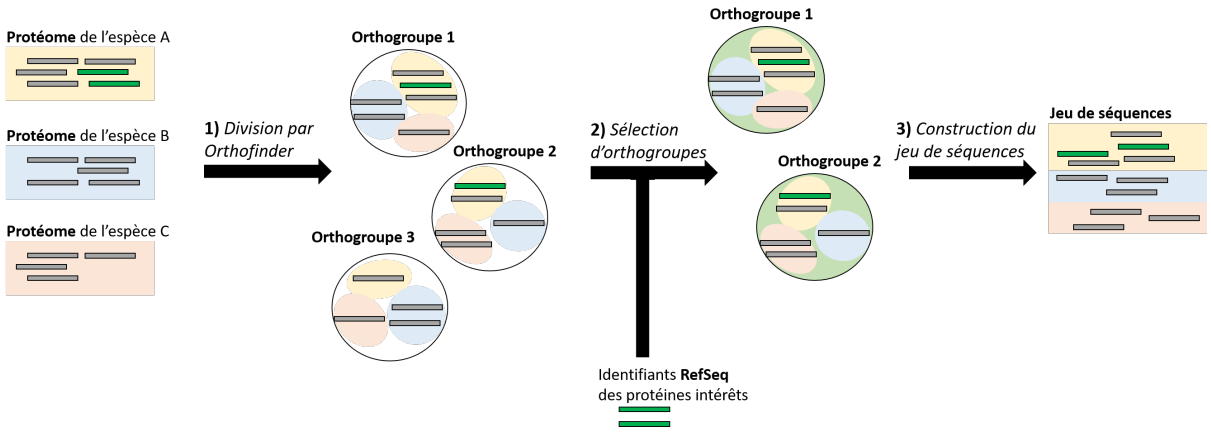

FIGURE 5.9 – **Construction d'un jeu de données de séquences à partir de protéomes**
1) Les protéomes sont divisés en *orthogroupes* avec `Orthofinder`, 2) sélection des *orthogroupes* contenant au moins l'une des protéines d'intérêt, 3) Le jeu de données de séquences est construit en agrégeant les séquences des protéines présentes dans les *orthogroupes* sélectionnés.

5.2.3 Filtrage des isoformes et sélection d'une séquence représentative

Le *clustering* avec `OrthoFinder` porte sur les protéines des protéomes, lesquelles incluent les différentes isoformes d'un même gène. Les protéines susceptibles d'être regroupées dans un même *orthogroupe* sont d'une part les protéines homologues (orthologues et paralogues). Ayant une histoire commune, elles conservent des ressemblances. Ce sont d'autre part des isoformes de chaque gène, chacune étant différente des autres en termes de composition en exons. Pour estimer l'histoire évolutive des ADAMTS-TSL à l'échelle des gènes seuls, les homologues nous intéressent, pas l'ensemble des isoformes. Au sein d'un même orthogroupe se mêlent orthologues, paralogues et isoformes, et dans une combinaison donnée protéome/orthogroupe se regroupent les paralogues d'une espèce et leurs isoformes. Il est alors question de regrouper et de filtrer les isoformes de manière à ne conserver qu'une séquence par gène. Dans le but de filtrer les isoformes, nous avons iden-

tifié deux problèmes : 1) l'identification des isoformes d'un même gène, et 2) la représentation des exons d'un gène, considérant ses isoformes.

5.2.3.1 Regroupement des produits d'un gène

Dans le but de filtrer les isoformes des gènes ADAMTS-TSL chez nos 9 espèces, nous cherchons ici à regrouper les protéines issues de l'expression d'un même gène. Pour ceci, nous avons opté pour une stratégie d'alignement avec le génome. Retrouver le locus génomique dont chaque protéine provient (i.e., reporter une séquence de protéine à une coordonnée d'un génome, Figure 5.10), nous permet de regrouper les isoformes d'un gène : les isoformes d'un même gène partagent par définition des exons communs.

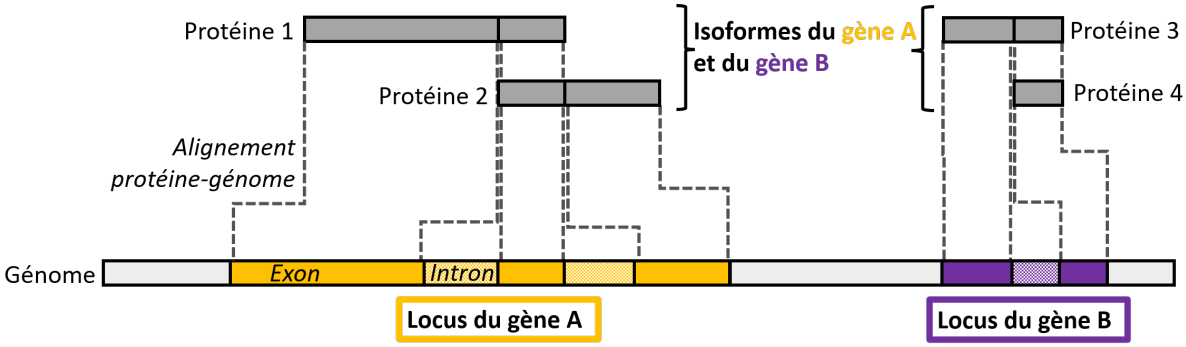

FIGURE 5.10 – **Alignement génomique des séquences de protéine et estimation de leurs loci génomiques.**
Les protéines 1, 2, 3 et 4 peuvent être repositionnées grâce aux annotations d'exons ou par alignement sur le génome (i.e., alignement génomique). La protéine 1 et la protéine 2 sont issues d'un même locus sur le même brin et sont donc deux isoformes d'un même gène.

Afin de positionner efficacement nos séquences protéiques sur les génomes de leur espèce, nous avons utilisé deux méthodes d'alignements distinctes, que nous appellerons *alignement-spaln* et *alignement-gff*. La différence majeure entre les deux méthodes d'alignement est la quantité d'*a priori* et d'information qu'elles utilisent. Les deux méthodes d'alignement utilisent comme entrée une protéine et le génome de l'espèce dont la protéine provient, et permettent d'obtenir le locus génomique de cette protéine. Un locus génomique est décrit par un chromosome/scaffold, le brin ainsi que les positions de début et de fin (Figure 5.11).

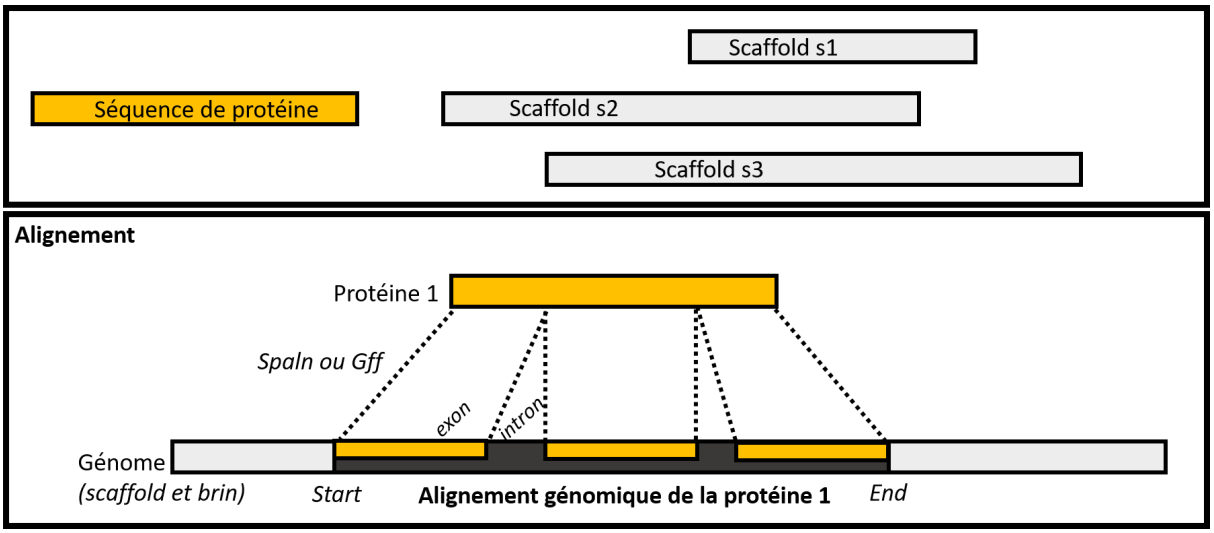

FIGURE 5.11 – **Alignement génomique d'une séquence de protéine sur un génome.**
La séquence d'une protéine peut être positionnée sur un génome (ici composé de trois *scaffold*), que ce soit par alignement sans *a priori* (`Spaln`), ou sur la base de ses annotations (`Gff`). Les positions d'une protéine sont décrites par un scaffold, un brin, une position de départ, une position de fin.

Alignement-spaln : basé sur la création d'alignements protéine-ADN par le logiciel *spaln* [IG12]. `Spaln` permet d'aligner une séquence de protéine avec un génome (séquence ADN), tout en considérant le passage d'un alphabet à un autre (code génétique) ainsi que la présence de jonctions d'exons. L'*alignement-spaln* ne nécessite ainsi aucune information supplémentaire et permet d'obtenir sans aucun *a priori* le locus d'une protéine. Cependant, il est nécessaire d'indexer en amont le génome, de manière à pouvoir effectuer l'alignement. Cette indexation ainsi que l'alignement peuvent être extrêmement coûteux en temps de calcul et en espace mémoire.

Alignement-gff : basé sur les annotations génomiques disponibles. Les annotations génomiques présentes dans un fichier GFF permettent de localiser directement le loci d'une protéine issue du protéome du même assemblage. Il suffit de chercher l'identifiant RefSeq de notre protéine dans les annotations GFF, et d'en extraire les informations de positions associées. Cette méthode est rapide, ne nécessite aucun calibrage et repose sur les connaissances antérieures de l'assemblage, cependant elle nécessite un fichier d'annotations GFF.

Après avoir positionné toutes les protéines sur leur génome, il est question de regrouper les protéines isoformes issues d'un même gène. Pour ceci, nous nous basons sur nos alignements génomiques (qu'ils soient réalisés avec la méthode spaln ou GFF). Nous regroupons pour un même gène les séquences des protéines dont les coordonnées génomiques sont : 1) sur la même séquence génomique (e.g., *scaffold*), 2) sur le même brin, 3) dont les coordonnées se chevauchent sur un minimum d'un acide aminé. La Figure 5.12 illustre un exemple d'alignement génomique de quatre protéines d'*H. sapiens*, quatre alignements qui permettent de regrouper quatre protéines en deux gènes paralogues distincts.

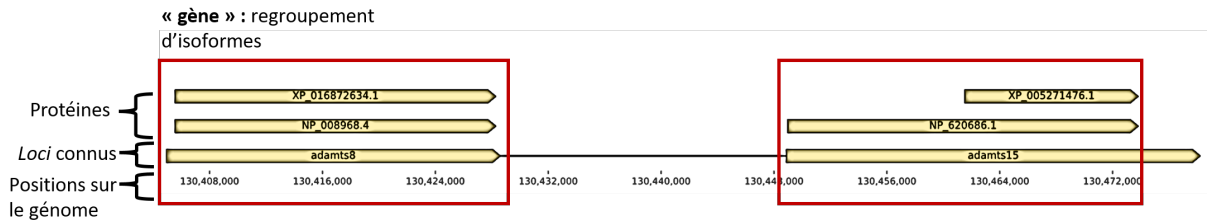

FIGURE 5.12 – **Exemple d'alignements de protéines sur le génome.**
Quatre séquences de protéines ADAMTS-TSL sont alignées avec Spaln sur la séquence génomique (*scaffold*) NC_000011.10 d'*H. sapiens*. Les *loci* connus des gènes ADAMTS-TSL sont indiqués en guise de références (i.e., ADAMTS-8, ADAMTS-15). L'alignement génomique nous permet de regrouper les séquences des isoformes par gènes (représentés ici par des rectangles rouges).

5.2.3.2 Séquence représentative et graphe de segments de gènes

Le regroupement des isoformes par gène permet de considérer l'information de séquences de l'ensemble des isoformes connus d'un gène. La variété des isoformes disponibles pour un gène pose le problème de savoir traiter cette information dans le but d'estimer une phylogénie des ADAMTS-TSL. En effet, considérer les différences entre les isoformes d'un gène ne s'interprète pas en termes de signal phylogénétique, mais d'une expression différente. De plus, les annotations fonctionnelles sont pour l'instant associées au gène, sans être spécifiques à une isoforme particulière, ce qui limite l'intérêt d'utiliser toutes les isoformes d'un gène. Il est donc question de choisir une séquence isoforme par gène dans le but d'inférer une phylogénie des gènes ADAMTS-TSL. Nous avons besoin de choisir des isoformes qui sont comparables (i.e., isoformes homologues). Cependant, identifier les isoformes homologues est un problème encore ouvert [GBB22], et identifier des isoformes homologues dans le cadre des gènes ADAMTS-TSL de 9 espèces serait un problème difficile. Toutefois, nous voulons éviter de comparer des isoformes qui ne sont pas comparables. Par exemple, nous souhaiterions éviter d'inférer une phylogénie avec une isoforme d'ADAMTS courte, ne possédant par exemple pas de domaine catalytique (e.g., l'isoforme X4 de ADAMTS-13, `Q76LX8-4` [Mar+95]), qui biaiserait la phylogénie obtenue par la ressemblance (non historique) de cette isoforme avec une séquence ADAMTSL sans domaine catalytique. Nous cherchons alors à sélectionner une isoforme représentative de son gène, tout en considérant les exons qui y seraient absents.

Nous avons ici fait le choix de sélectionner l'isoforme la plus longue comme séquence représentative de son gène, posant ainsi l'hypothèse que les différentes séquences représentatives sélectionnées correspondent à des séquences d'isoformes homologues. Cependant, sélectionner une isoforme pose le problème des exons spécifiques aux autres isoformes. Dans le but d'estimer l'information présente dans les isoformes non utilisées (exons absents de la séquence représentative), nous avons mis au point une représentation des isoformes d'un gène sous forme de graphe, que l'on nommera *graphe de segments* (Figure 5.13). Les *graphes de segments* se basent sur le principe des graphes d'épissages [Lac+08 ; Ber+14], tout en reposant sur une segmentation en blocs par le programme `Paloma` [CK05] (utilisant un alignement multiple local et partiel, décrit plus en détail en section 3.1.3). Le but est de représenter les régions identiques de protéines isoformes sous forme de segments uniques (i.e., les exons d'un même gène présents dans plusieurs isoformes), de manière à factoriser l'information des séquences isoformes d'un gène. Pour ceci, les séquences des isoformes d'un gène sont segmentées en blocs représentant les similarités locales partagées

par les isoformes. Chaque bloc est ensuite représenté par une séquence, appelée segment (i.e., suite d'exons consécutifs présents dans toutes les isoformes possédant ce bloc). Finalement, l'enchaînement de ces segments constitue le *graphe de segments* et représente l'information codante du gène. Notons que le graphe est orienté de l'extrémité N-terminale à l'extrémité C-terminale des protéines isoformes (Figure 5.13).

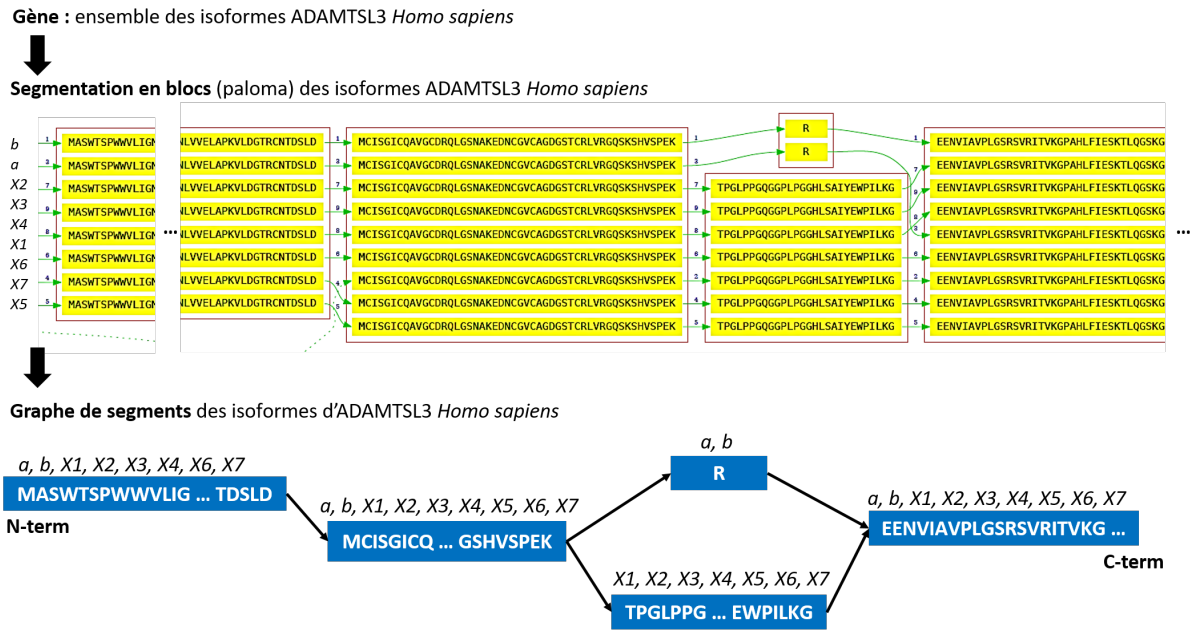

FIGURE 5.13 – **Construction d'un *graphe de segments* d'isoformes avec le programme `Paloma`.**
Le graphe de segment (partie inférieure) représente les segments de séquences partagés par les isoformes du gène. Un segment est représenté par une boîte bleue, sa séquence est écrite à l'intérieur, les isoformes le possédant sont indiquées au-dessus (e.g., *a, b, X1*). Les flèches représentent l'enchaînement des segments au sein des isoformes. Dans cet exemple, il y a un saut d'exon dans les isoformes *a, b*, de l'exon commun aux isoformes *X1, X2, X3, X4, X5, X6, X7*.

Nous avons implémenté cette méthode, comprenant la génération et la visualisation des *graphes de segments*. La Figure 5.14 est un exemple de visualisation que nous avons généré, ici pour le *graphe de segments* des isoformes du gène ADAMTS-like 4 humain.

La représentation en graphe du contenu en séquences des isoformes (i.e., *graphe de segments*), nous permet de considérer la totalité de l'information des isoformes d'un gène, de visualiser l'isoforme la plus longue, ainsi que les segments absents de certaines isoformes (saut d'exon). Nous utilisons ces *graphes de segments* dans le cadre de la quantification et la représentation des segments d'isoformes absents de la séquence représentative (en rouge sur la Figure 5.15).

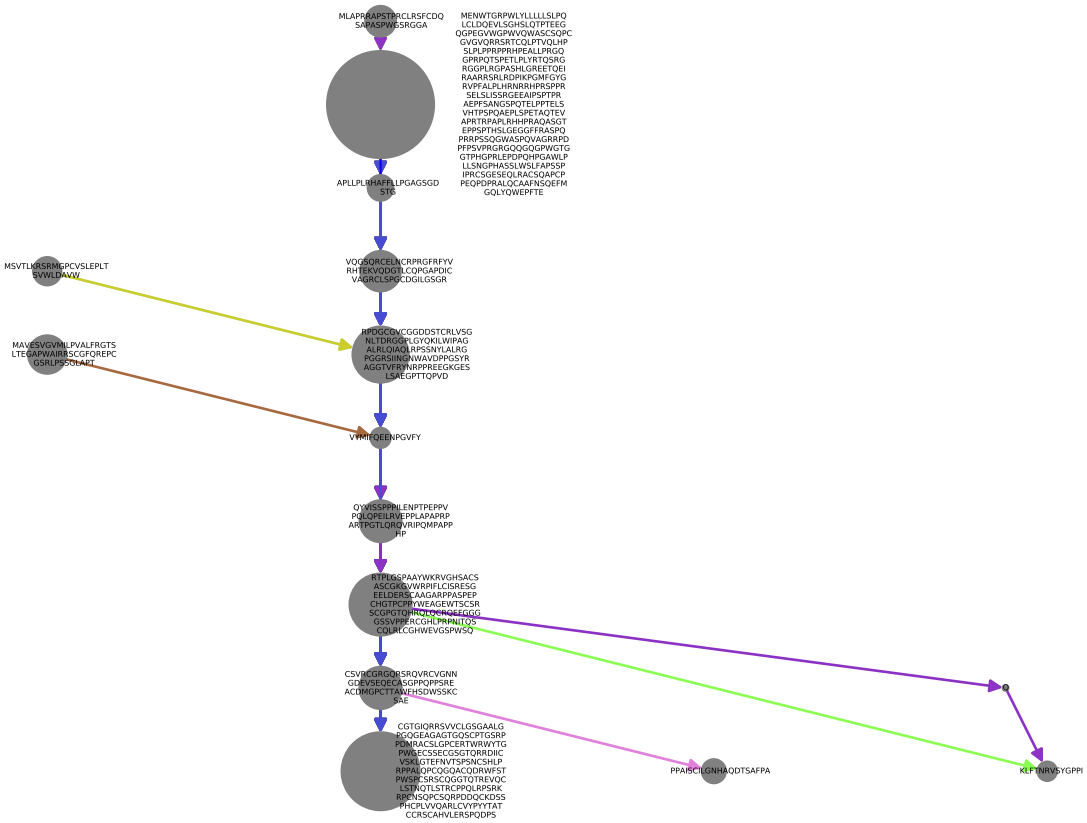

FIGURE 5.14 – **Le *graphe de segments* des isoformes du gène ADAMTS-like 4 humain.** Généré automatiquement à partir de la liste des séquences protéiques associées au gène ADAMTS-like 4 humain. Les séquences issues d'un orthogroupe ont été segmentées avec le programme `Paloma` (paramètres t3, m1, M11, c) afin de construire le *graphe de segments*. Chaque nœud correspond à un segment, soit une suite d'exons consécutifs dans plusieurs isoformes, dont la taille dépend de la longueur de la séquence. Un arc indique l'enchaînement de deux segments dans une isoforme donnée. La couleur des arcs indique un enchaînement propre à une isoforme donnée.

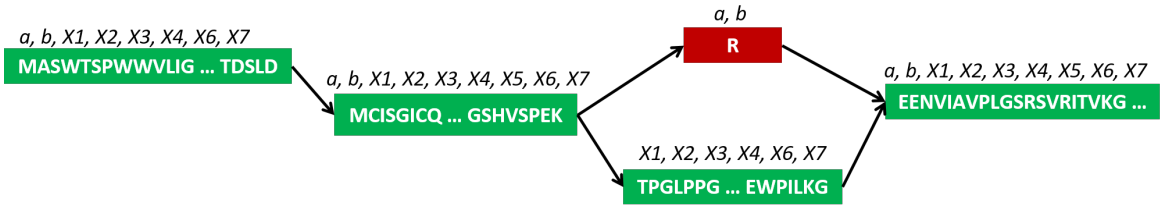

FIGURE 5.15 – **Sélection de l'isoforme la plus longue (vert) comme représentante du gène.** Les segments en vert appartiennent à la séquence de l'isoforme la plus longue (X1), considérée comme représentative du gène. Le segment en rouge n'est pas présent dans X1 et correspond à un saut d'exon qui ne sera pas considéré.

Finalement, aligner les séquences de protéines d'une même espèce sur le génome correspondant nous permet de regrouper les séquences d'isoformes d'un même gène. Dans le but d'inférer une phylogénie des gènes ADAMTS-TSL considérant une isoforme par gène, nous sélectionnons la séquence de l'isoforme la plus longue comme séquence représentative du gène, tout en visualisant les exons qui n'y seraient pas présents (qui ne seront pas considérés dans les analyses ultérieures). L'objectif est de maximiser la quantité de séquences analysées dans les étapes ultérieures.

5.2.4 Construction du jeu de données de référence

Nous présentons le jeu de données de séquences construit avec le protocole présenté dans les deux sections précédentes (5.2.2 et 5.2.3). Nous avons précédemment sélectionné 9 espèces de bilatériens (5.2.1) possédant un protéome et un génome annoté de qualité (Tableau 5.2), comme représentantes de l'histoire évolutive de la famille ADAMTS-TSL dans la lignée humaine, allant de l'ancêtre des *Bilateria* à *Homo sapiens*. Après sélection d'un assemblage pour chaque espèce (Tableau 5.2), nous avons récupéré le protéome correspondant au format fasta et les annotations génomiques associées au format GFF (*Gene Feature Format*).

TABLE 5.2 – **Les 9 espèces, leur taxid et assemblage sélectionné.**

Espèce	Taxid	Assemblage
Homo sapiens	9606	GCF_000001405.39
Mus musculus	10090	GCF_000001635.27
Bos taurus	9913	GCF_002263795.1
Gallus gallus	9031	GCF_016699485.2
Xenopus tropicalis	8364	GCF_000004195.4
Danio rerio	7955	GCF_000002035.6
Ciona intestinalis	7719	GCF_000224145.3
Drosophila melanogaster	7227	GCF_000001215.4
Caenorhabditis elegans	6239	GCF_000002985.6

En utilisant les 9 protéomes comme données d'entrée pour l'outil `Orthofinder` [EK19] (v2.5.2, paramètres par défaut), nous avons regroupé leurs séquences en 31,229 *orthogroupes*. Parmi eux, 22 *orthogroupes* comprennent au moins un des 159 identifiants RefSeq des ADAMTS (102), ADAMTSL (57) connus chez l'humain, 17 *orthogroupes* correspondent aux protéines ADAMTS et 5 aux protéines ADAMTSL. Quatre gènes ADAM

humains (ADAM10, ADAM17, ADAM9 et ADAM12), représentatifs des protéines de la famille ADAM ont été utilisés comme groupe externe, les 7 identifiants RefSeq y correspondant nous ont permis d'identifier 2 *orthogroupes* les contenant. La totalité de ces 24 *orthogroupes* correspond à 708 séquences (386 ADAMTS, 226 ADAMTSL, 96 ADAM). Nous avons nommé ce jeu de données *dataset-708* (décrit dans le Tableau 5.3). Après regroupement des isoformes par gène (méthode *alignement-gff*) et sélection d'une séquence représentative pour chacun des gènes, nous avons construit le *dataset-214*, qui contient 214 séquences de protéines représentatives de 214 gènes, issues de 9 espèces choisies : 173 ADAMTS-TSL (125 ADAMTS, 48 ADAMTSL), plus 41 ADAM (Table 5.3). Il faut noter que, par construction, les 26 gènes ADAMTS-TSL humains sont représentés dans le *dataset-214*.

TABLE 5.3 – **Détail du jeu de données des 9 espèces.**
Le dataset-708 contient toutes les séquences des 24 orthogroupes sélectionnés (plusieurs isoformes par gènes).
Le dataset-214 contient une séquence représentative par gène.

	ADAMTS	ADAMTSL	ADAM	Total				
Orthogroupes	17	5	2	24				
	Dataset-708				Dataset-214			
Espèces	ADAMTS	ADAMTSL	ADAM	Total	ADAMTS	ADAMTSL	ADAM	Total
H. sapiens	102	68	20	190	19	7	4	30
M. musculus	106	37	11	154	19	7	4	30
B. taurus	42	32	7	81	19	7	4	30
G. gallus	55	38	17	110	18	6	5	29
X. tropicalis	31	16	9	56	19	8	6	33
D. rerio	34	19	14	67	22	7	8	37
C. intestinalis	9	11	5	25	6	4	4	14
D. melanogaster	5	3	9	17	2	1	3	6
C. elegans	2	2	4	8	1	1	3	5
Total	386	226	96	708	125	48	41	214

En considérant les relations d'homologies entre toutes les protéines de 9 espèces de *Bilateria*, nous avons construit le jeu de séquences *dataset-214*, comprenant une liste exhaustive de protéines homologues aux 26 ADAMTS-TSL humaines et les 4 ADAM utilisées comme groupe externe. Ce jeu de séquences *dataset-214* sera utilisé comme jeu de référence. Il contient 214 séquences de protéines représentatives de 214 gènes qui seront

analysées pour déterminer la phylogénie des ADAMTS-TSL (Section suivante, 5.3), les contenus en modules conservés (Chapitre 6) et les phénotypes ancestraux (Chapitre 7).

5.3 Inférence et contrôle de la phylogénie de référence des ADAMTS-TSL

Bien que l'évolution des ADAMTS ait été étudiée à plusieurs reprises (voir Section 4.1), aucune étude ne propose de phylogénie associant ADAMTS et ADAMTSL au sein d'un unique arbre. C'est pourquoi nous avons construit un jeu de séquences homologues aux 26 ADAMTS-TSL humaines, le plus exhaustif possible, et dont les 9 espèces sélectionnées permettent de considérer huit ancêtres durant l'évolution des ADAMTS-TSL humaines depuis l'ancêtre *Bilateria*. Notre objectif est alors d'inférer une phylogénie de référence des 214 gènes ADAMTS-TSL (Section 5.3.1) du jeu de séquences *dataset-214*, d'en tester la robustesse (Section 5.3.2) et d'effectuer différents contrôles sur les données et méthodes utilisées (Section 5.3.3).

5.3.1 Phylogénie de référence des ADAMTS-TSL

Afin d'inférer un unique arbre phylogénétique que nous utiliserons comme référence dans la suite de la thèse, nous avons utilisé 214 séquences homologues chez 9 espèces *dataset-214*, comprenant des ADAMTS-TSL ainsi que des ADAM comme groupe externe. Nous allons présenter la méthodologie choisie (Section 5.3.1.1) ainsi que l'arbre de référence obtenu (Section 5.3.1.2) que nous confronterons à la littérature (Section 5.3.1.3).

5.3.1.1 Inférence de l'arbre de référence des gènes ADAMTS-TSL

Ce paragraphe traite de l'inférence d'un arbre de gènes ADAMTS-TSL de référence, à partir de 214 séquences de protéines de 214 gènes homologues (orthologues et paralogues chez 9 espèces) et de l'arbre des 9 espèces. La famille ADAMTS-TSL est composée de différents sous-groupes de gènes ayant des compositions en domaines différentes, à l'intérieur desquels les gènes sont très similaires entre eux (e.g., hyalectanases, procollagenases, ADAMTS-like). Cette caractéristique nous a incité à utiliser l'outil `PASTA` [Mir+15] pour construire l'alignement multiple (MSA, pour ***M****ultiple* ***S****equence* ***A****lignement*) de nos 214 séquences de protéines. En effet, `PASTA` repose sur un principe itératif d'alignements de sous-ensemble de séquences, permettant de construire un MSA robuste pour un grand

nombre de séquences divergentes. Dans le cas des ADAMTS-TSL, effectuer un nettoyage [2] de l'alignement ferait perdre la quantité d'information importante qu'apporte la disparité des compositions en domaines des différentes protéines.

Nous inférons avec l'outil `RaxML` [Sta14] une première phylogénie à partir de la totalité du MSA de `PASTA`, sans aucun nettoyage ou sélection. Nous utilisons ensuite le programme `TreeFix` [Wu+13] pour réconcilier ce premier arbre de 214 gènes avec l'arbre phylogénétique des 9 espèces. `TreeFix` va modifier la topologie de l'arbre des gènes de manière à la rendre cohérente avec celle de l'arbre des espèces, tout en gardant une vraisemblance quasi maximale, ce qui permet de corriger des nœuds peu soutenus correspondants à des erreurs stochastiques. Le nouvel arbre des gènes résultant de `TreeFix` possède une topologie prenant en compte l'alignement `PASTA` ainsi que l'évolution des espèces dont les gènes proviennent. Cependant, ce nouvel arbre des gènes ne contient plus d'information comme les longueurs de branche et les supports (e.g., *bootstrap*). C'est pourquoi nous recalculons ensuite les longueurs des branches et les supports (approximation Bayes) avec l'outil `PhyML` [Gui+10] considérant l'alignement `PASTA` et la topologie `TreeFix` obtenue. Finalement, l'arbre sera enraciné manuellement avec le groupe externe constitué de séquences de protéines ADAM, le tout aboutissant à notre arbre de référence.

5.3.1.2 Arbre de référence ADAMTS-TSL

Appliquer ce protocole aux 214 séquences du *dataset-214* nous a permis d'inférer l'arbre de référence des gènes ADAMTS-TSL (Figure 5.16). Cet arbre de référence nous servira de base pour la suite de nos inférences (compositions en modules et phénotypes ancestraux). Cependant, nous allons effectuer divers contrôles dans la suite de ce chapitre, afin d'identifier quels regroupements dans la phylogénie sont robustes au protocole de reconstruction.

L'analyse des supports de branches Bayes (Figure 5.17) montre que la majorité des nœuds sont fortement soutenus. Seuls les nœuds les plus profonds, proche de la racine et certains nœuds proches des feuilles, sont plus faiblement soutenus. Pour ce qui est des nœuds proches des feuilles, leur faible support Bayes peut s'expliquer par la forte similarité de séquences entre les ADAMTS-TSL des espèces proches et du manque de signal phylogénétique que cela implique. Ces nœuds en particulier sont « corrigés » avec

2. Un nettoyage d'alignement multiple consiste à enlever toutes les colonnes peu informatives (e.g., avec un grand nombre d'*indels*), et à ne sélectionner que les colonnes les plus informatives (possédant un résidu aligné pour toutes les séquences). Historiquement, `Gblock` [Cas00] et `Trimal` [CSG09] sont des outils très utilisés à cette fin.

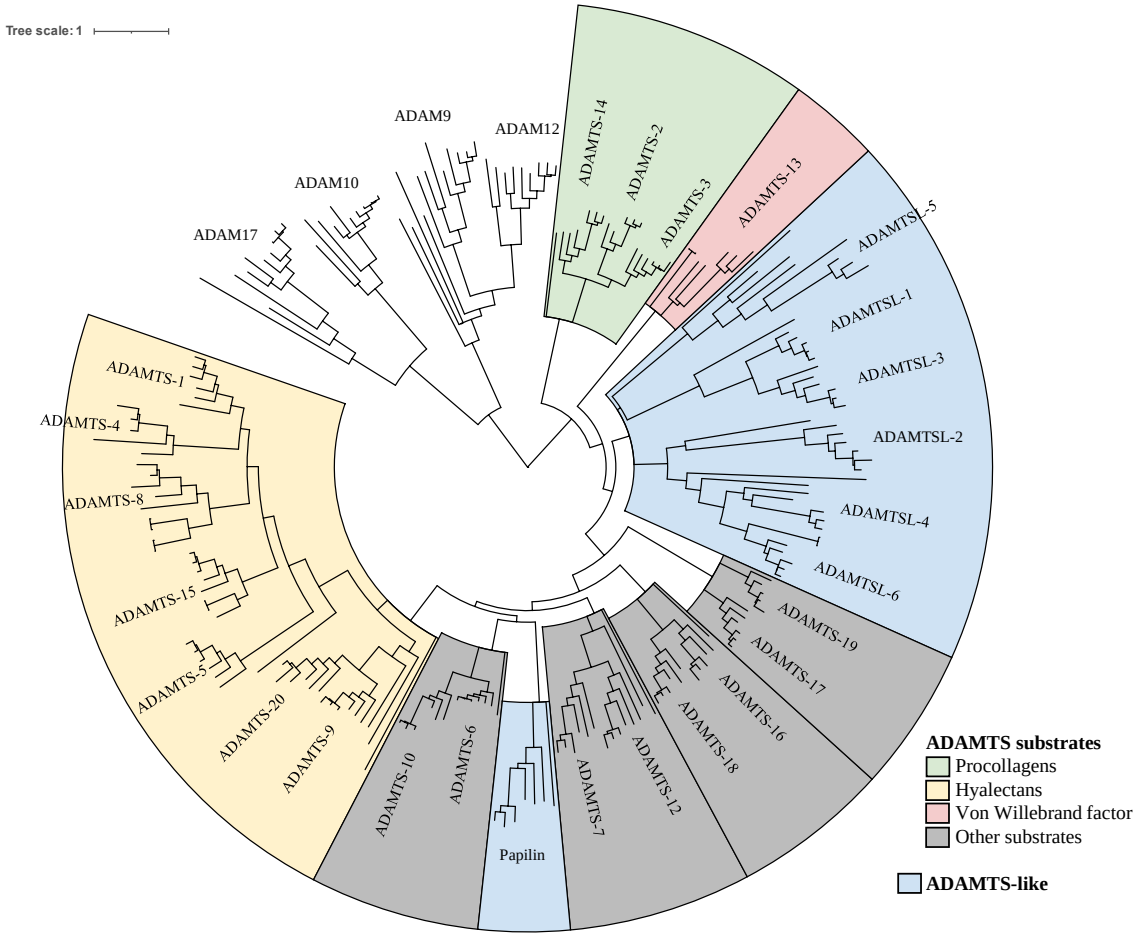

FIGURE 5.16 – **Arbre phylogénétique de référence de 214 ADAMTS-TSL.**
Les 214 séquences ont été alignées avec `PASTA` et la phylogénie a été inférée avec `RaxML`. La topologie de l'arbre a ensuite été corrigée avec `TreeFix`, sachant l'arbre des espèces. Les 41 séquences d'ADAM ont été utilisées pour enraciner l'arbre. Les longueurs des branches ont été calculées avec `PhyML` sachant la topologie finale et l'alignement `PASTA`. Les différents clades sont colorés en fonction des groupes fonctionnels connus.

`TreeFix` de manière à privilégier une résolution de l'arbre compatible avec l'arbre des espèces. Les branches plus internes qui sont faiblement soutenues, suggèrent une difficulté à inférer avec assurance les relations les plus anciennes entre les sous-groupes de protéines. Nous étudierons plus en détail la robustesse du scénario évolutif proposé ici, dans les Sections 5.3.2 et 5.3.3.

5.3.1.3 Comparaison avec la littérature

Nous allons maintenant comparer l'arbre de référence obtenu avec ceux des études précédentes qui proposaient des phylogénies ADAMTS-TSL [Hux+05 ; Nic+05 ; Hux+07 ; Bru+15]. L'utilisation des protéines ADAM comme groupe externe nous permet de proposer une phylogénie enracinée de 173 ADAMTS-TSL issues de 9 espèces de *Bilateria* qui est consistante avec les conclusions principales de ces études, en particulier sur l'expansion paralogue de la famille au cours de l'expansion des vertébrés, résultant d'événements de duplications de génomes complets [Bru+15]. De façon similaire à l'étude utilisant une séquence ADAMTS de *Porifera* [Bru+15] comme groupe externe, notre arbre enraciné identifie les procollagénases (ADAMTS-2, -3, -14) comme le premier groupe de gènes à diverger au sein des ADAMTS-TSL. Comme décrit par [Bru+15], nous avons identifié une monophylie des procollagénases (ADAMTS-2, -3, -14) ainsi qu'une monophylie des hyalectanases (ADAMTS-1, -4, -5, -8, -9, -15 and -20), avec des relations similaires à l'intérieur de ces clades. De plus, nous avons également identifié une monophylie des COMP protéases (ADAMTS-7, -12) et des autres paires d'enzymes connues comme des paralogues proches (i.e. ADAMTS-6 et ADAMTS-10, ADAMTS-16 et ADAMTS-18, ADAMTS-17 et ADAMTS-19). Un résultat notable est la monophylie des ADAMTSL, excluant la Papilin (pourtant généralement considérée comme une ADAMTSL), qui se regroupe en un clade distinct avec ADAMTS-6 et ADAMTS-10. Nos résultats proposent donc une origine des ADAMTSL et de la Papilin au sein des ADAMTS par perte de leur domaine catalytique. Ce qui nous mène à poser l'hypothèse novatrice que les ADAMTSL seraient des pseudoenzymes dérivées des ADAMTS, comme c'est le cas pour différentes autres familles d'enzymes [MFE17].

Finalement, la phylogénie de référence des gènes ADAMTS-TSL que nous avons inférée représente l'évolution de la famille chez 9 espèces de bilatériens en incluant les 26 gènes ADAMTS-TSL humains et confirme les observations déjà publiées tout en apportant de nouvelles hypothèses sur l'origine des ADAMTSL.

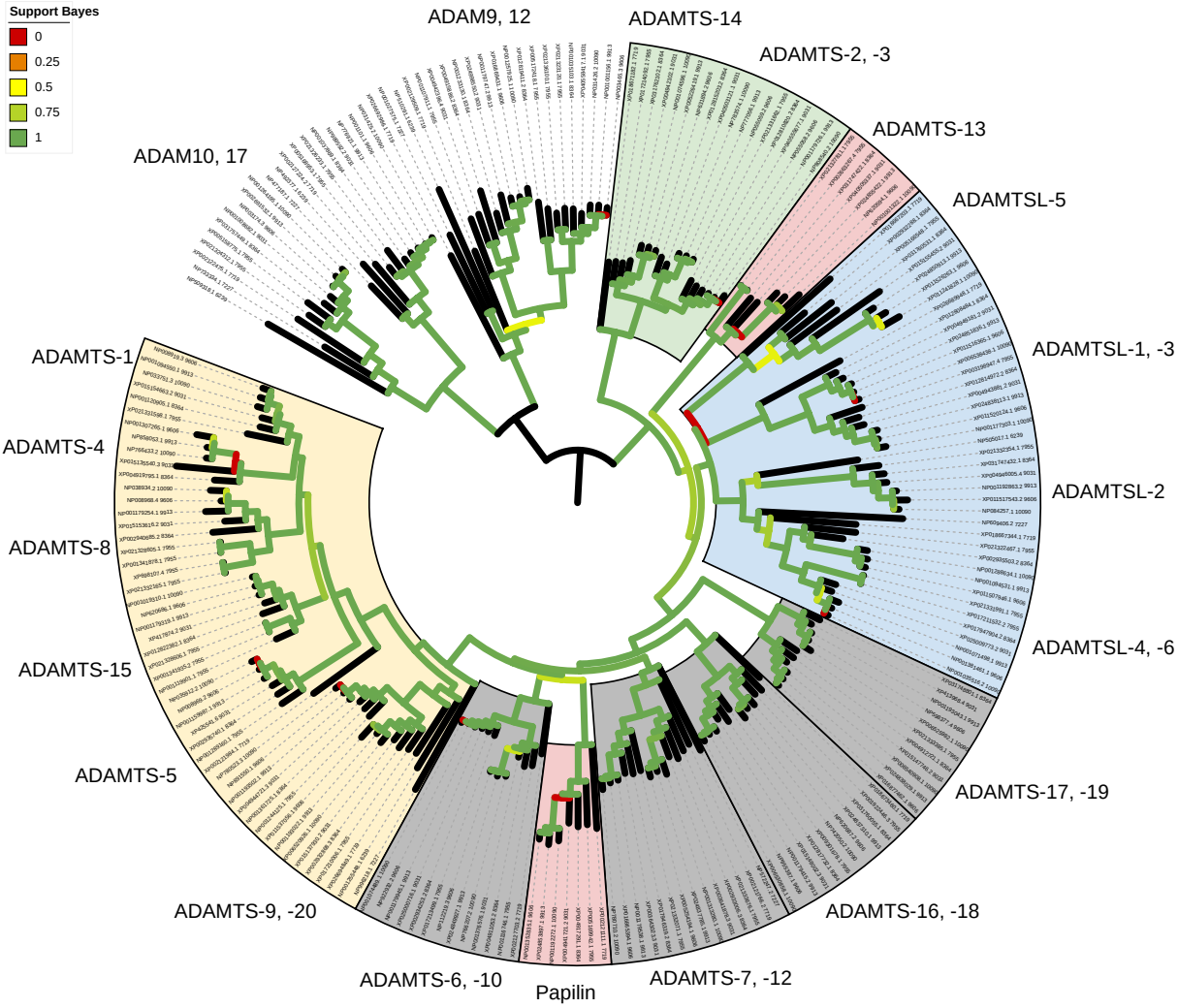

FIGURE 5.17 – **Supports Bayes de l'arbre phylogénétique de référence ADAMTS-TSL.** Les 214 séquences ont été alignées avec `PASTA` et la phylogénie a été inférée avec `RaxML`. La topologie de l'arbre a ensuite été corrigée avec `TreeFix`, sachant l'arbre des espèces. Les 41 séquences d'ADAM ont été utilisées pour enraciner l'arbre (**Lien Itol**). Les longueurs de branches ainsi que les valeurs de supports Bayes ont été calculées avec `PhyML` sachant la topologie finale. Les valeurs de supports Bayes des branches sont indiquées par leurs couleurs, allant de 0 en rouge, à 100 en vert.

5.3.2 Robustesse de la phylogénie de référence à l'échantillonnage taxonomique

Un arbre phylogénétique peut être sujet à des biais dus aux séquences sélectionnées pour l'inférer. Nous avons effectué différents rééchantillonnages du jeu de séquences utilisé, dans le but d'estimer la robustesse de l'arbre de référence et des ancêtres pour lesquels nous établirons des conclusions dans les chapitres suivants. Nous commencerons par présenter la méthode de rééchantillonnage taxonomique utilisée (Section 5.3.2.1), avant d'en présenter une application sur un jeu de séquences alternatif (Section 5.3.2.2). Pour finir, nous effectuerons différents rééchantillonnages du *dataset-214* afin d'estimer la robustesse des différents clades de l'arbre de référence (Section 5.3.2.3).

5.3.2.1 Support des bipartitions et robustesse à l'échantillonnage taxonomique

Les bipartitions, définies comme des ensembles de feuilles disjointes dans un arbre phylogénétique, peuvent varier quand le protocole de reconstruction dont il provient varie. Cependant, deux arbres phylogénétiques aux topologies différentes peuvent tout de même partager des bipartitions (Figure 5.18). Au cours de cette section, nous allons nous intéresser aux bipartitions les plus présentes au sein d'un jeu d'arbres d'ADAMTS-TSL, c'est-à-dire aux bipartitions les plus robustes, dans le but d'identifier les regroupements récurrents d'ADAMTS-TSL. En ce sens, une bipartition robuste est définie comme insensible à la méthode de reconstruction utilisée.

Afin d'estimer la robustesse de notre phylogénie de référence, nous avons opté pour un rééchantillonnage taxonomique d'un jeu de séquences (présenté en Figure 5.19). Pour ceci, nous utilisons deux jeux de séquences, le premier étant celui qui nous intéresse (séquences du jeu de référence), le second comprenant des séquences homologues appartenant à des espèces non considérées dans le premier jeu (séquences alternatives). Les deux jeux de séquences d'intérêts et de séquences alternatives sont ainsi mutuellement exclusifs. Le principe est le suivant : nous sélectionnons n séquences du jeu de séquences alternatives, que l'on ajoute au jeu de données d'intérêt. Nous obtenons ainsi un jeu de séquences comprenant toutes les séquences d'intérêts ainsi que quelques séquences alternatives sélectionnées aléatoirement. Ensuite, nous réalisons l'alignement multiple de ce jeu de séquences rééchantillonné, avant d'en inférer la phylogénie. En itérant i fois ce rééchantillonnage, nous obtenons i alignements multiples et i phylogénies différentes. Parmi

ces i arbres phylogénétiques, la présence de séquences alternatives différentes peut faire varier les alignements, et donc les topologies. Ce qui nous permet d'étudier l'occurrence et la robustesse des bipartitions possibles pour nos séquences d'intérêts. Une bipartition présente dans un grand nombre d'arbres issus du rééchantillonnage sera considérée comme robuste au choix des données.

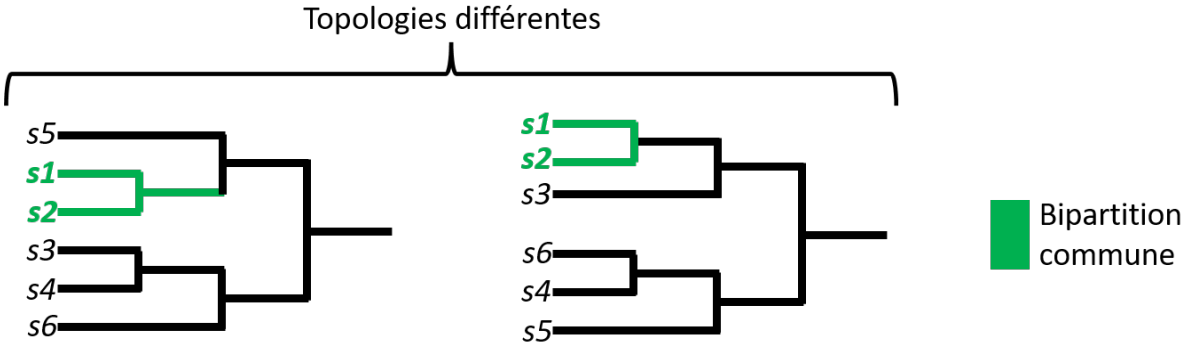

FIGURE 5.18 – **Bipartition partagée par deux arbres de topologies différentes.**
Deux arbres phylogénétiques des mêmes séquences peuvent posséder des topologies différentes. Cette différence de topologie peut provenir de différences dans les méthodes utilisées pour les inférer. Les variations entre ces topologies sont ainsi dues à une sensibilité à la méthode utilisée et se traduisent par des bipartitions non partagées par les différents arbres (clade s3+s4 présent sur le premier arbre, absent sur le deuxième). Cependant, certaines bipartitions (comme s1+s2) sont communes aux différentes topologies, signe de regroupements de séquences robustes à la méthode utilisée (en vert).

Dans les paragraphes suivants, nous estimerons les bipartitions robustes, sur la base de différents rééchantillonnages et de différents jeux de données de séquences ADAMTS-TSL, en appliquant le principe décrit ci-dessus. Le but est d'identifier des regroupements d'ADAMTS-TSL robustes au choix des séquences analysées.

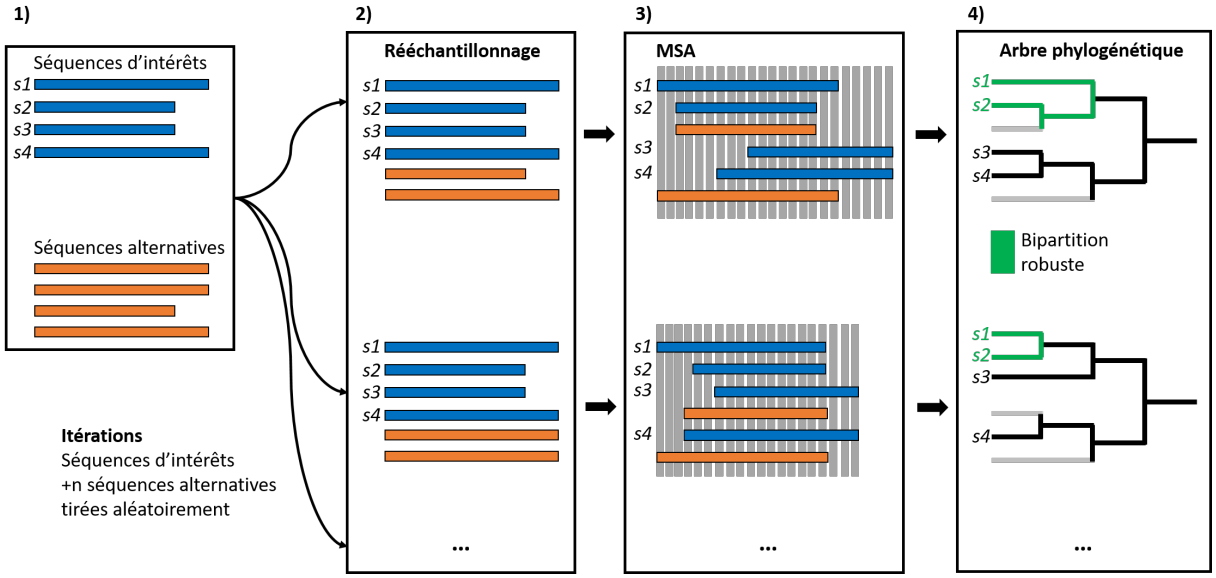

FIGURE 5.19 – **Effet du rééchantillonnage taxonomique sur la topologie.**
1) Deux jeux de séquences : séquences d'intérêts (s1, s2, s3, s4) et séquences alternatives. Tirage aléatoire de n séquences du jeu de séquences alternatives qui sont ajoutées aux séquences d'intérêts. **2)** Jeu de séquences contenant les séquences d'intérêts et n séquences alternatives. **3)** Alignement multiple du jeu de séquences rééchantillonnées. **4)** Inférence de la phylogénie. Itérer i fois ce rééchantillonnage permet d'obtenir i alignements multiples et i phylogénies différentes. Un groupe (bipartition) sera dit robuste s'il n'est pas sensible au choix des séquences considérées, comme c'est le cas ici pour s1+s2. Le score de robustesse est estimé en calculant la fréquence des différentes bipartitions dans les i phylogénies.

5.3.2.2 Exemple sur 8 espèces

Nous avons étudié la robustesse des bipartitions en utilisant des méthodes de calcul d'arbres consensus. Un arbre consensus synthétise l'information d'un grand nombre d'arbres construits avec les mêmes séquences. Nous utilisons ici une méthode d'arbre consensus bien spécifique, appelé *consensus Adams* [Ada72 ; JLS17] [3]. Le principe de l'arbre *consensus Adams* et de considérer l'information de tous les arbres, une bipartition de l'arbre *consensus Adams* sera nécessairement présente dans tous les arbres considérés en entrée. Ainsi, les sous-arbres et bipartitions présents dans l'arbre consensus sont des bipartitions robustes qui sont présentes dans la totalité des arbres. Appliquer un tel consensus à des arbres issus de rééchantillonnage nous permet d'identifier des sous-arbres ou des regroupements de gènes ADAMTS-TSL dont l'estimation est insensible aux variations dues aux séquences considérées en entrée.

Nous avons ici utilisé un jeu de données de 177 ADAMTS-TSL (8 espèces [4]), que nous avons rééchantillonné 100 fois, en ajoutant à chaque fois 20 séquences sélectionnées aléatoirement d'un jeu de données ADAMTS-TSL alternatif (espèces non considérées parmi les 8 [5]). Ceci nous a permis d'obtenir 100 arbres dont nous avons ensuite retiré les feuilles correspondantes aux séquences alternatives, pour enfin construire l'arbre *consensus Adams* de 100 arbres à 177 feuilles (Figure 5.20). Une grande partie des bipartitions ne sont pas robustes et remontent directement à la racine de l'arbre, indiquant que les relations entre certains groupes sont sensibles au choix initial des séquences. Cependant, trois sous-arbres se distinguent (en vert, en bleu et en rouge). Le sous-arbre vert contient ADAMTS-1, -4, -5, -8, -9, -15, -20, ce qui correspond au groupe fonctionnel des hyalectanases. Le groupe rouge contient ADAMTS-2, -3, -14, ce qui correspond au groupe fonctionnel des procollagénases. Le groupe bleu contient les couples ADAMTS-7/ADAMTS-12, ADAMTS-16/ADAMTS-18 et ADAMTS-6/ADAMTS-10. Ces trois groupes peuvent être considérés avec confiance : ils sont robustes à l'échantillonnage taxonomique et soutenus par le fait qu'une même fonction est partagée par les séquences du sous-groupe (i.e., partage un même ligand). Pour ce qui est des bipartitions qui remontent jusqu'à la racine (en noir), elles contiennent les différentes ADAMTS-like, ADAMTS-13, -17, -19, indiquant que la position de ces gènes au sein d'une phylogénie des ADAMTS-TSL n'est pas robuste à

3. *Adams* est le créateur de la méthode, et n'a aucun rapport avec la famille ADAM, ADAMTS-TSL
4. *Homo sapiens, Mus musculus, Bos taurus, Gallus gallus, Xenopus tropicalis, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans*
5. *Rattus norvegicus, Xenopus laevis, Canis lupus familiaris, Felis catus, Anolis carolinensis, Pan troglodytes*

l'échantillonnage taxonomique et donc incertaine.

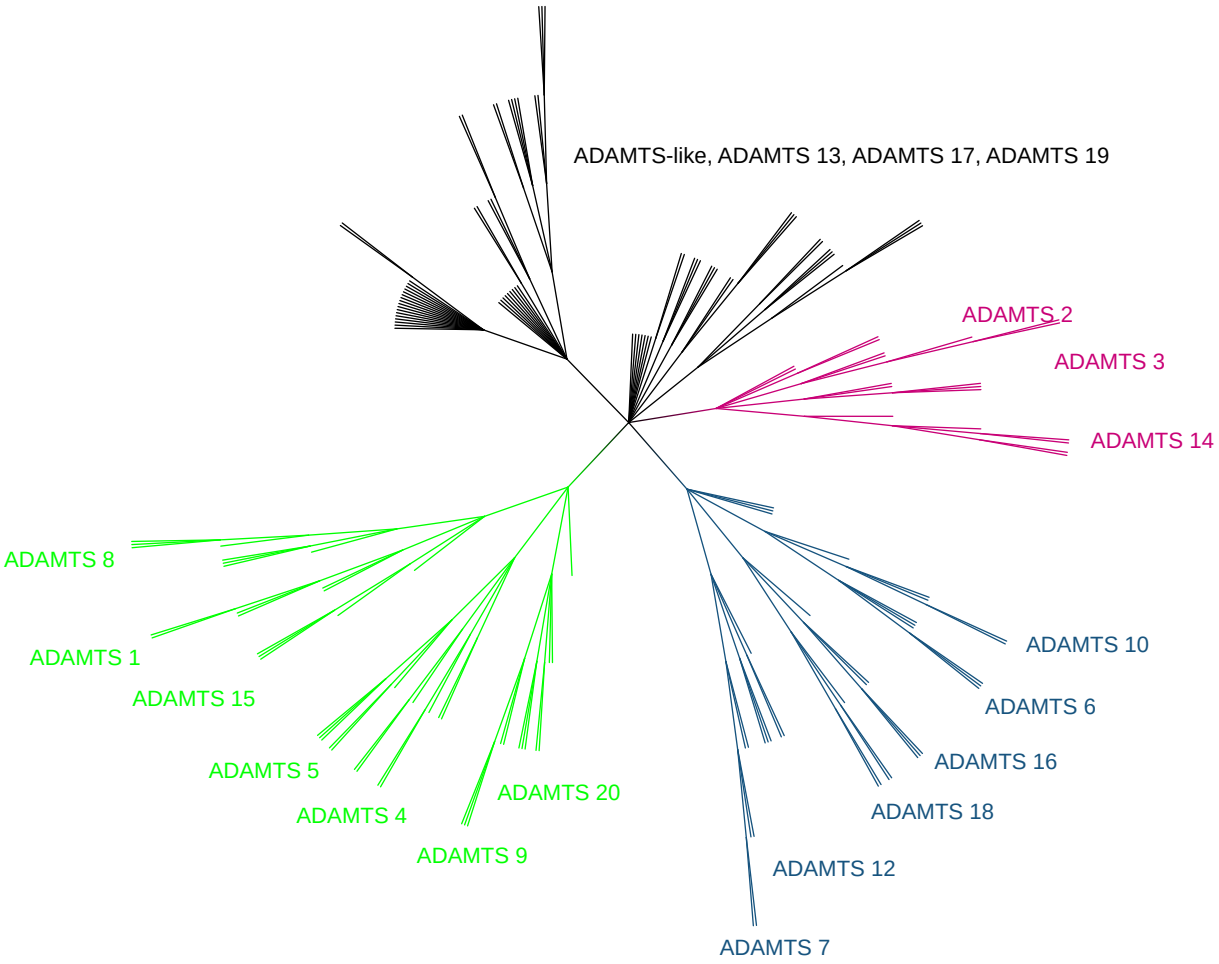

FIGURE 5.20 – **Arbre consensus *Adams* du rééchantillonnage taxonomique de 177 ADAMTS-TSL.**
Les 177 protéines homologues ADAMTS-TSL issues de 8 espèces (*Homo sapiens, Mus musculus, Bos taurus, Gallus gallus, Xenopus tropicalis, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans*) ont été rééchantionnées 100 fois par ajout de 20 séquences sélectionnées aléatoirement dans un jeu de données ADAMTS-TSL alternatif (contenant d'autres espèces). Les 100 arbres sont inférés avec `PhyML` après alignement avec `Muscle`. Les 20 séquences qui avaient été ajoutées sont ensuite retirées de l'arbre. Un arbre *consensus Adams* est ensuite construit à partir de ces 100 arbres représentant les groupes présents dans tous les arbres. L'arbre consensus n'est pas enraciné. Les trois sous-arbres robustes sont colorés en vert, rouge et bleu.

5.3.2.3 Robustesse de l'arbre de référence

Notre arbre de référence sera utilisé comme base pour les analyses de modules et phénotypes ancestraux. Il est question ici d'appliquer différents rééchantillonnages du jeu de données initial, dans le but d'identifier les nœuds et bipartitions robustes et ceux qui ne le sont pas. L'étude complète inclut deux types de rééchantillonnage : un du groupe externe, l'autre du groupe interne (cf. Annexe 10.1 pour plus de détails).

Le rééchantillonnage du groupe externe utilise différents sous-groupes de protéines ADAM (groupe externe). Nous avons construit six jeux de données différents, chacun incluant nos 125 ADAMTS et 48 ADAMTSL ainsi qu'un sous-groupe parmi les 41 séquences du groupe externe ADAM. Ces séquences incluent soit 1) ADAM9, soit 2) ADAM12, soit 3) ADAM17, soit 4) ADAM10, soit 5) ADAM9 + ADAM12 ainsi que 6) ADAM17 + ADAM10.

Le rééchantillonnage du groupe interne utilise des séquences d'ADAMTS-TSL d'espèces différentes des 9 considérées dans le jeu de référence. Nous avons construit un jeu de séquences ADAMTS-TSL alternatif, composé de 255 séquences connues (Uniprot canonique) de six espèces non considérées dans notre jeu de séquences de référence (*Rattus norvegicus, Xenopus laevis, Canis lupus familiaris, Felis catus, Anolis carolinensis, Pan troglodytes*). Ensuite, nous avons construit 20 jeux de séquences différents, chacun contenant nos 173 séquences ADAMTS-TSL de références, ainsi que 10 séquences choisies aléatoirement dans le jeu de séquences alternatif. Pour tous ces jeux, les 41 séquences ADAM du jeu de référence sont considérées comme groupe extérieur.

Pour tous ces jeux de séquences, l'alignement de séquences est calculé avec `PASTA` et la phylogénie est obtenue avec `RaxML`. Ce qui nous intéresse ici, ce sont les informations que donnent ces deux rééchantillonnages indépendants sur notre arbre de référence. En effet, chaque nœud de l'arbre de référence est associé avec une valeur de robustesse, représentant le nombre de fois où sa bipartition associée est retrouvée dans les phylogénies issues des rééchantillonnages taxonomiques. Nous obtenons ainsi des valeurs de robustesse issues du rééchantillonnage du groupe externe (Figure 5.21) et du rééchantillonnage du groupe interne (Figure 5.22). Ces deux rééchantillonnages sont tous deux en accord avec nos observations préalables : les nœuds récents (proches des feuilles) sont fortement robustes (en vert) contrairement aux nœuds les plus ancestraux (proches de la racine) qui ne sont que faiblement robustes (en rouge). En particulier, les monophylies des groupes hyalectanases et procollagénases sont estimées robustes aux rééchantillonnages, ainsi que les quatre couples : ADAMTS-6/-10, ADAMTS-7/-12 (COMP protéases), ADAMTS-16/-

18, ADAMTS-17/-19.

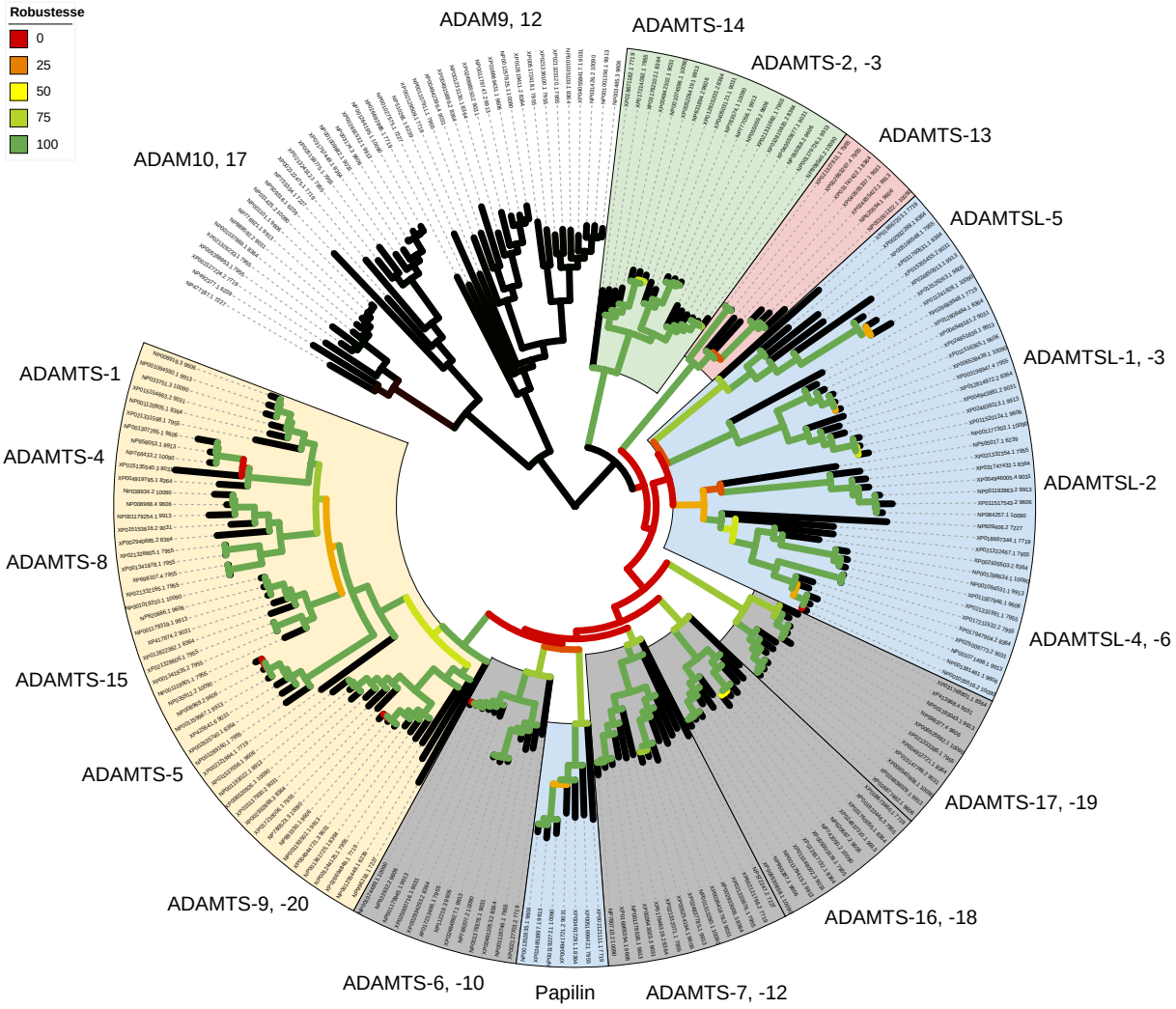

FIGURE 5.21 – **Arbre de référence avec valeurs de robustesse du rééchantillonnage du groupe externe.**
Le rééchantillonnage du groupe externe utilise différents sous-groupes de protéines ADAM (groupe externe). Nous avons construit 6 jeux de données différents, chacun incluant nos 125 ADAMTS et 48 ADAMTSL ainsi qu'un sous-groupe parmi les 41 séquences du groupe externe ADAM. Ces séquences incluent soit 1) ADAM9, soit 2) ADAM12, soit 3) ADAM17, soit 4) ADAM10, soit 5) ADAM9 + ADAM12 ainsi que 6) ADAM17 + ADAM10. Nous avons ensuite construit les MSA avec `PASTA` avant d'inférer la phylogénie pour chacun de ces jeux de données avec `RaxML`. Finalement, nous avons calculé les valeurs de robustesse des clades du groupe interne de l'arbre de référence (couleur des branches, allant de 0 en rouge, à 100 en vert), à partir des arbres rééchantillonnés. La valeur de robustesse d'un clade de l'arbre de référence représente la proportion d'arbres rééchantillonnés où la bipartition associée est présente.

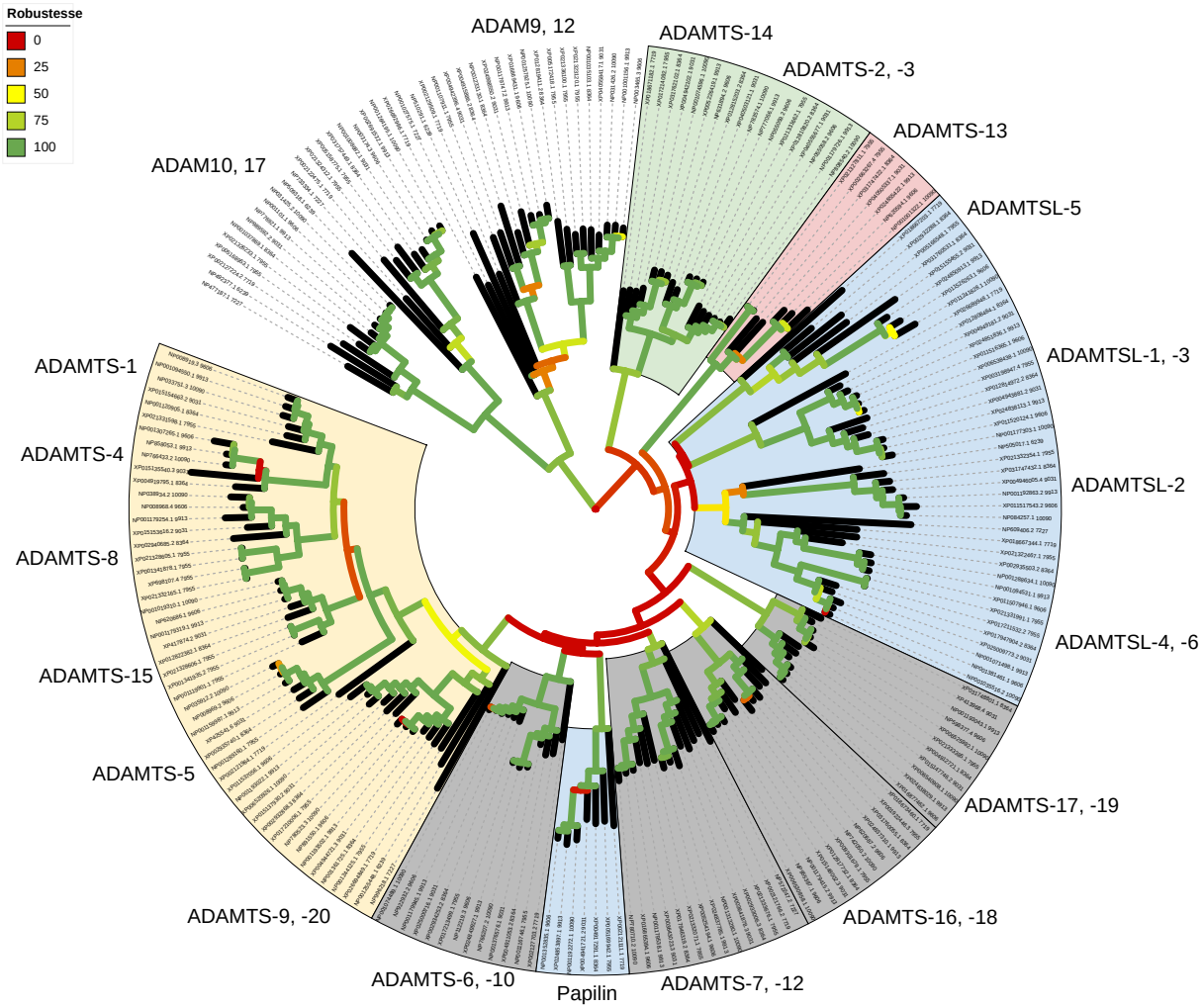

FIGURE 5.22 – **Arbre de référence avec valeurs de robustesse du rééchantillonnage du groupe interne.**
Le rééchantillonnage du groupe interne utilise des séquences d'ADAMTS-TSL de 6 espèces différentes. Nous avons construit un jeu de séquences ADAMTS-TSL alternatif, composé de 255 séquences connues (Uniprot canonique) de 6 espèces non considérées dans notre jeu de séquences de référence (*Rattus norvegicus, Xenopus laevis, Canis lupus familiaris, Felis catus, Anolis carolinensis, Pan troglodytes*). Ensuite, nous avons construit 20 jeux de séquences différents, chacun contenant nos 173 séquences ADAMTS-TSL de référence, ainsi que 10 séquences choisies aléatoirement dans le jeu de séquences alternatif. Nous avons ensuite construit les MSA avec `PASTA` avant d'inférer la phylogénie pour chacun de ces jeux de données avec `RaxML`. Finalement, nous avons calculé les valeurs de robustesse des clades de l'arbre de référence (couleur des branches, allant de 0 en rouge, à 100 en vert), à partir des arbres rééchantillonnés. La valeur de robustesse d'un clade de l'arbre de référence représente la proportion d'arbres rééchantillonnés où la bipartition associée est présente.

5.3.3 Contrôle des données et des méthodes exploitées pour calculer la phylogénie de référence

Afin de tester la robustesse de la phylogénie de référence à la méthode et aux données utilisées, nous avons également fait des contrôles en faisant varier deux éléments : 1) le choix du programme d'alignement multiple, et 2) le choix des gènes considérés en groupe externe.

5.3.3.1 Robustesse des phylogénies aux méthodes d'alignement

Généralement, pour inférer une phylogénie à partir de séquences, il faut : 1) aligner les séquences, 2) nettoyer (ou non) l'alignement et 3) inférer l'arbre à partir de l'alignement obtenu. Chacune de ces étapes est l'objet d'un choix, que ce soit le choix d'une méthode d'alignement multiple, le choix de nettoyer ou non l'alignement et le choix des méthodes d'inférence de l'arbre. En effet, inférer un arbre phylogénétique revient à appliquer un modèle d'évolution sur un alignement multiple (MSA), c'est pourquoi les différences entre deux alignements multiples ont un effet bien plus important que les paramétrages du modèle évolutif [OR06 ; Wan+11 ; Md +15 ; Ash+19]. Dans cette partie, nous allons nous focaliser sur les effets du choix d'un alignement multiple sur la topologie de l'arbre phylogénétique des protéines ADAMTS-TSL.

Pour un même jeu de séquences, l'utilisation d'alignements multiples différents peut être source d'arbres phylogénétiques aux topologies différentes. De manière à tester cette variation, nous avons aligné un même jeu de 341 séquences ADAMTS-TSL (décris en 5.1.3.1) avec six différents programmes d'alignement multiple (`Clustal Omega` [Sie+11], `Mafft` [Kat+02], `Muscle` [Edg04], `Tcoffe` [NHH00], `Prank` [Löy14], `Dialign` [Mor99 ; AYM13]), avant d'en inférer la phylogénie puis de comparer les topologies des arbres obtenus. Nous présentons ici les topologies des six arbres phylogénétiques calculés avec `PhyML` et nettoyés avec `Trimal` (Figure 5.23). Les sous-groupes fonctionnels sont identifiés par des couleurs différentes. Cette comparaison met en évidence un fort effet de l'alignement multiple sur les relations des ADAMTS-TSL dans une phylogénie. En accord avec nos observations antérieures, les groupes fonctionnels de gènes sont robustes, et nous retrouvons les monophylies des hyalectanases (jaune), procollagénases (vert) et des couples ADAMTS-6/-10, ADAMTS-7/-12, ADAMTS-16/-18, ADAMTS-17/-19 (gris). Cependant, les relations entre ces sous-groupes ne sont aucunement robustes à la méthode d'alignement multiple utilisée.

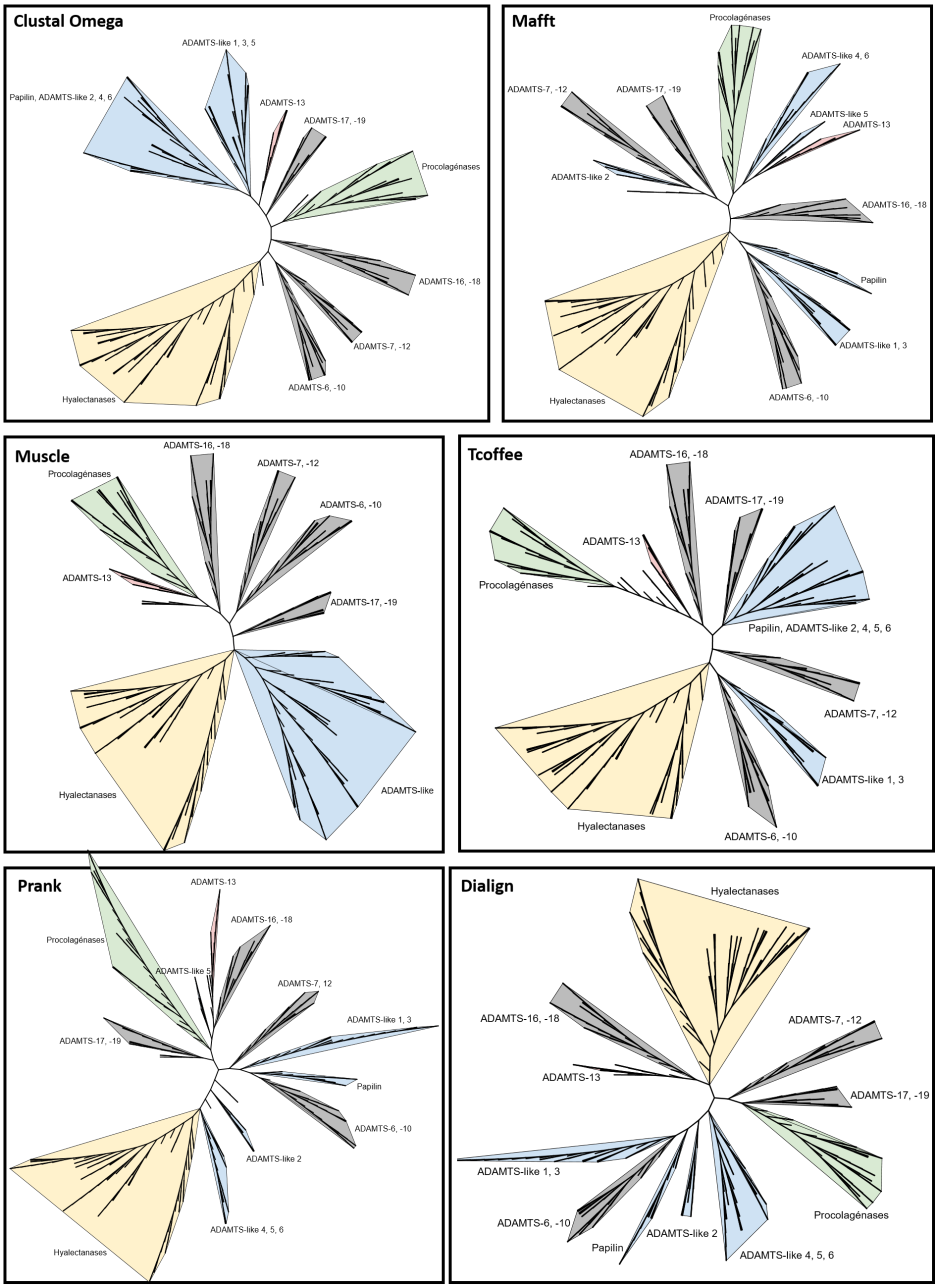

FIGURE 5.23 – **Comparaison de la topologie de l'arbre phylogénétique des ADAMTS-TSL (341 séquences) pour six méthodes d'alignements multiples différentes.**
341 séquences d'ADAMTS-TSL sont alignées avec six programmes d'alignement multiple différents (`Clustal Omega` [Sie+11] **Lien Itol**, `Mafft` [Kat+02] **Lien Itol**, `Muscle` [Edg04] **Lien Itol**, `TCoffe` [NHH00] **Lien Itol**, `Prank` [Löy14] **Lien Itol**, `Dialign` [Mor99 ; AYM13] **Lien Itol**). Chaque alignement obtenu est ensuite nettoyé avec `Trimal` (Paramètre Automated 1) [CSG09] et nous inférons leurs phylogénies avec l'outil `PhyML` [Gui+10]. Aucun arbre n'est enraciné, et les longueurs de branches ne sont pas considérées.

Nous montrons ici que le choix de la méthode de construction de l'alignement multiple utilisé pour l'inférence phylogénétique a un impact direct sur la topologie des arbres phylogénétiques, dans le cas des ADAMTS-TSL. La question que nous nous posons n'est pas de savoir quelle méthode d'alignement est la « meilleure », mais bien de rechercher les bipartitions et les regroupements communs à tous les alignements obtenus, c'est-à-dire les sous-ensembles de séquences dont l'origine commune est estimée de manière robuste à la méthode utilisée.

5.3.3.2 Sélection d'un groupe externe

La quasi-totalité des modèles phylogénétiques utilisés sont dits « temps réversibles », et les arbres produits sont non enracinés : il n'est pas possible d'identifier la racine (origine du temps) dans l'arbre obtenu. De manière classique, un groupe externe doit être utilisé pour enraciner l'arbre d'intérêt, ou groupe interne. Les groupes externes et internes ont une origine commune et seule la phylogénie du groupe interne est examinée. Dans le cadre de l'étude d'une famille de protéines, le choix d'un groupe externe n'est pas évident. Utiliser un groupe externe, revient à poser l'hypothèse de l'existence d'un ancêtre commun entre notre famille de protéines ADAMTS-TSL (le groupe interne) et le groupe externe, soit quatre gènes de la famille ADAM dans notre phylogénie de référence.

Trois groupes externes potentiels : choisir un groupe externe dans le cadre de l'étude d'une famille de protéines est parfois complexe. Il est question de choisir des protéines possédant une origine évolutive commune avec les ADAMTS-TSL et si possible qui ont peu divergé. Nous avons identifié trois familles de Metzincin ayant une origine évolutive avec les ADAMTS-TSL (Figure 5.24) : 1) les *Snake Venom MetalloProteases* (SVMP), 2) les *Matrix MetalloProteinases* (MMP) et les 3) *A Disintegrin And Metalloproteases* (ADAM) [Hux+07]. Afin de comparer ces potentiels groupes externes, nous avons inféré différentes phylogénies d'ADAMTS-TSL avec chacun d'eux.

Choix du groupe externe : de manière à étudier l'effet de ces trois groupes externes sur une phylogénie des ADAMTS-TSL, nous avons construit trois jeux de données, contenant 341 séquences d'ADAMTS-TSL (19 espèces, décrit en 5.1.3.1), ainsi que des séquences représentant les différents groupes externes : MMP, SVMP et ADAM. Les trois jeux de séquences ont ensuite été alignés avec le logiciel `Muscle`, avant d'en inférer la phylogénie avec le programme `PhyML`. Nous avons ensuite comparé les trois arbres obtenus, particulièrement la manière dont les trois groupes externes enracinent le groupe interne (Figures 5.25-5.26-5.27).

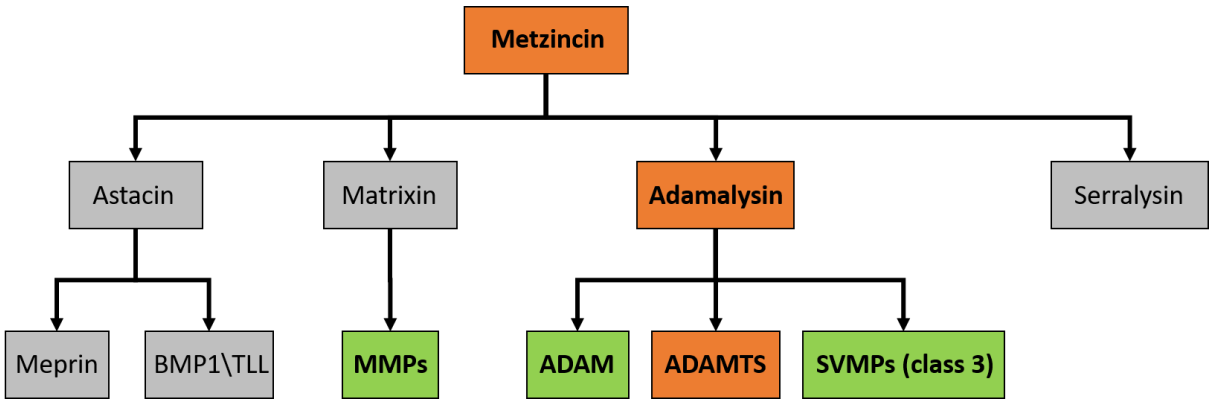

FIGURE 5.24 – **Classification des protéines de la famille des Metzincins**, adapté de [Hux+07]. Les SVMP, les MMP et les ADAM (vert) constituent des groupes externes potentiels pour enraciner l'arbre phylogénétique des protéines ADAMTS-TSL.

Notre groupe externe doit remplir deux critères : 1) le groupe externe doit être monophylétique, 2) de ne pas avoir trop divergé (i.e., ne pas avoir des branches trop longues, notamment la branche basale). Un groupe externe ayant trop divergé fait perdre du signal phylogénétique et nous risquons des erreurs dites d'attraction des longues branches [Ber05]. Nos trois groupes externes candidats sont monophylétiques, remplissant le premier critère. Ils peuvent ainsi être utilisés pour enraciner un arbre ADAMTS-TSL. Le groupe externe MMP présente des longueurs de branches bien plus importantes que le reste de l'arbre (Figure 5.26), indiquant une évolution rapide [6]. À l'opposé, le groupe externe SVMP présente des longueurs de branches bien inférieures au reste de l'arbre (Figure 5.25), indiquant une évolution plus lente au sein du groupe externe. Pour ces deux groupes externes (MMP, SVMP) nous observons de plus, de très longues branches séparant les groupes externes et internes (supérieur à un remplacement par site). Les ADAM, utilisées en groupe extérieur, sont plus proches des ADAMTS-TSL (Figure 5.27), et la branche séparant les groupes externe et interne est plus courte (inférieur à un remplacement par site). Ce critère nous a conduits à préférer les ADAM plutôt que les MMP ou les SVMP, pour enraciner le groupe interne ADAMTS-TSL, car les risques d'artefacts dus à l'attraction des longues branches pourraient être réduits. Nous observons également que le choix du groupe externe à un effet important sur la topologie du groupe interne, en particulier sur les nœuds internes profonds du groupe interne et son enracinement. De manière importante, les groupes hyalectanases et procollagénases, ainsi que les ADAMTS-7, -12 qui

6. La longueur d'une branche est l'espérance du nombre de substitutions par sites

seront examinés en détail dans la suite de la thèse (Chapitre 8), apparaissent robustes au choix du groupe extérieur.

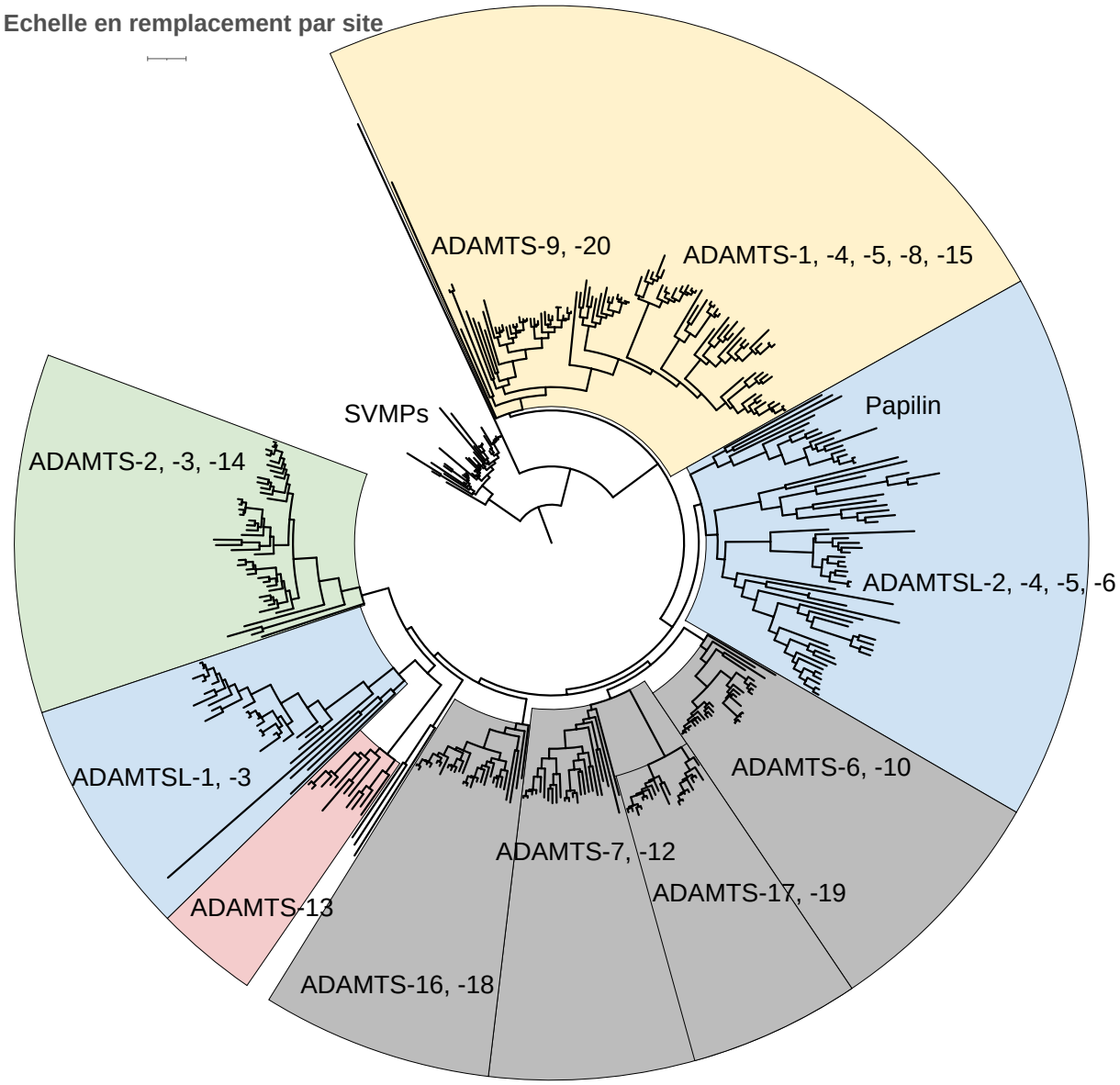

FIGURE 5.25 – **Arbre phylogénétique des 341 ADAMTS-TSL avec 38 SVMP comme groupe externe.**
Les 379 (341 ADAMTS-TSL + 38 SVMP) séquences sont alignées avec `Muscle` et la phylogénie est inférée en utilisant `PhyML`.

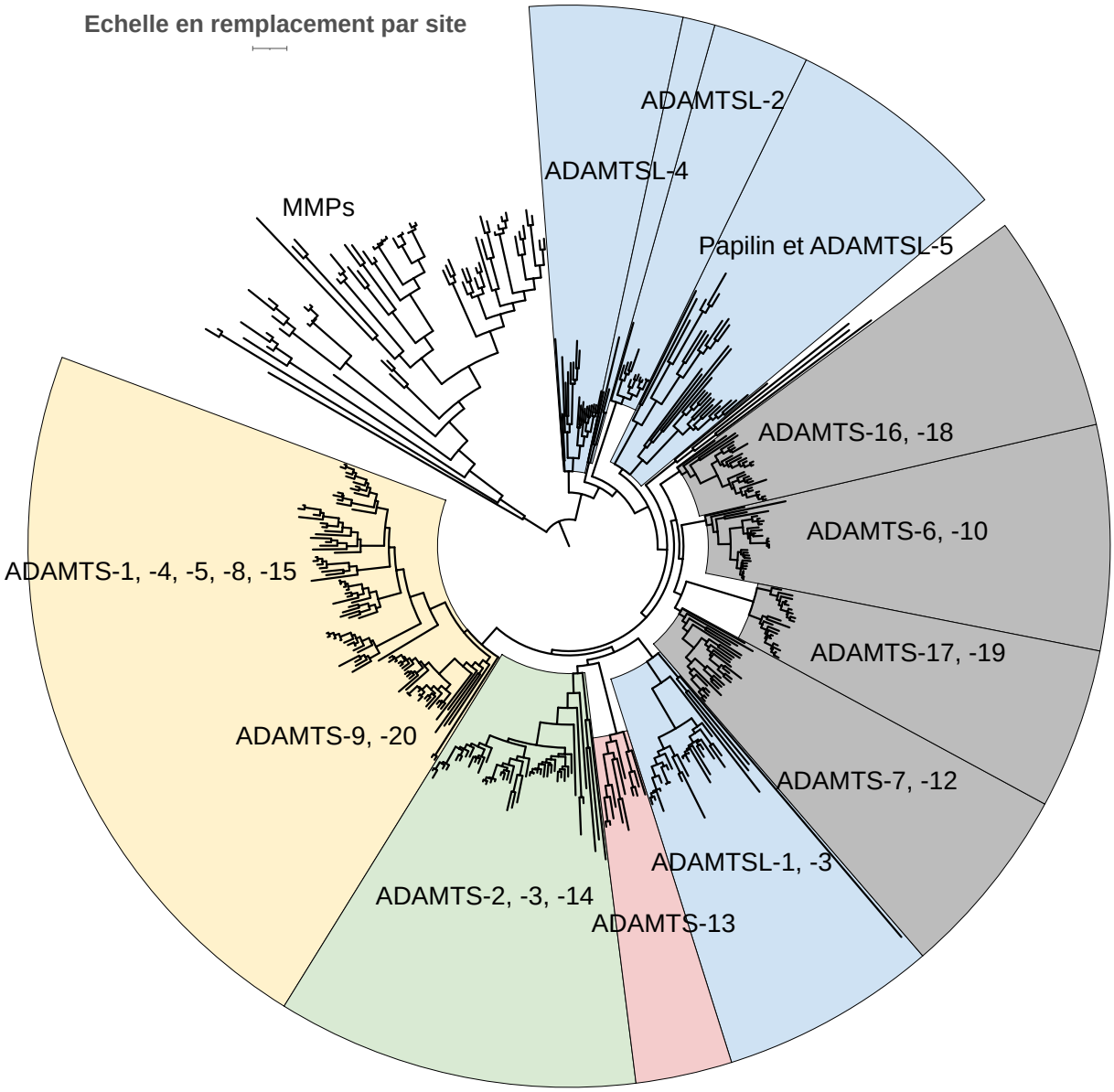

FIGURE 5.26 – **Arbre phylogénétique des 341 ADAMTS-TSL avec 65 MMP comme groupe externe.**
Les 406 (341 ADAMTS-TSL + 65 MMP) séquences sont alignées avec `Muscle` et la phylogénie est inférée en utilisant `PhyML`.

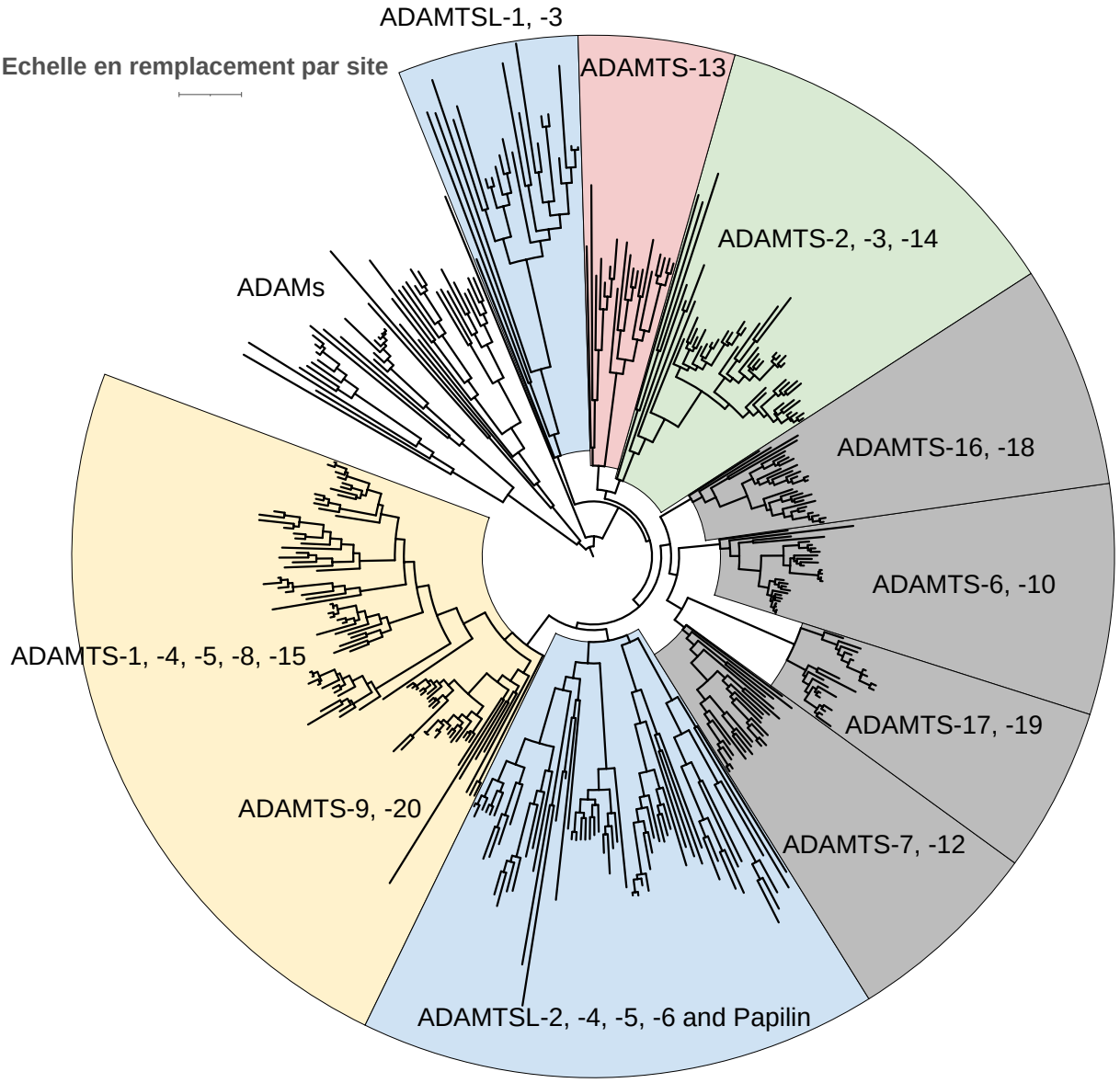

FIGURE 5.27 – **Arbre phylogénétique des 341 ADAMTS-TSL avec 41 ADAM comme groupe externe.**
Les 382 (341 ADAMTS-TSL + 41 ADAM) séquences sont alignées avec `Muscle` et la phylogénie est inférée en utilisant `PhyML`.

Conclusion

Au cours de ce chapitre, nous nous sommes intéressés aux séquences ADAMTS-TSL, afin de créer un jeu de données de séquences considérant homologues (orthologues, paralogues) et isoformes. Notre première contribution est la création d'un jeu de données d'ADAMTS-TSL de 9 espèces, contenant 214 gènes, leurs séquences représentatives, et les 708 séquences de leurs isoformes. Notre deuxième contribution est l'inférence de l'arbre phylogénétique de référence des 214 séquences représentatives ADAMTS-TSL. Cet arbre phylogénétique de référence nous servira donc de modèle pour toutes les analyses qui vont suivre. En effet, c'est sur cet arbre de référence que nous allons inférer l'évolution des modules et des interactions protéine-protéine, dans le but d'associer modules et fonctions. Par son importance cruciale dans ce projet, nous avons étudié la robustesse des différentes bipartitions de cet arbre, afin de ne pas obtenir uniquement un arbre référence comme modèle, mais aussi un arbre de référence dont nous savons quels nœuds sont robustes au choix des séquences, et donc propres à inférer des prédictions, et lesquels sont trop peu soutenus pour être considérés. Les groupes des hyalectanases, des procollagénases et des ADAMTS-7, -12, qui seront examinés en détails, apparaissent en particulier robustes au choix du protocole utilisé pour estimer la phylogénie des ADAMTS-TSL.

Chapitre 6

Évolution des compositions en modules des ADAMTS-TSL

Dans le cas des protéines ADAMTS-TSL qui sont multidomaines et multifonctionnelles, les compositions des séquences en domaines ne suffisent pas à discriminer fonctionnellement les différentes copies paralogues. Notre objectif est de chercher des signatures fonctionnelles plus fines et associées à des différences de fonction des ADAMTS-TSL humaines. Nous présentons ici une approche basée sur la recherche de conservations locales sans *a priori*, que nous appellerons *modules*. Au cours de ce chapitre (Figure 6.1), nous allons décrire ce qu'est un module (Section 6.1), la segmentation des séquences ADAMTS-TSL en une décomposition de modules (Section 6.2), avant de considérer la phylogénie propre à chaque module (Section 6.3) afin d'inférer l'histoire évolutive des compositions en modules le long de notre arbre des gènes ADAMTS-TSL de référence (Section 6.4). Ceci nous permettra d'étudier les gains et les pertes de modules au cours de l'évolution des gènes ADAMTS-TSL, et de décrire des groupes de modules gagnés de manière simultanée au cours de l'évolution, qui seront par la suite associés à des phénotypes (Chapitre 8).

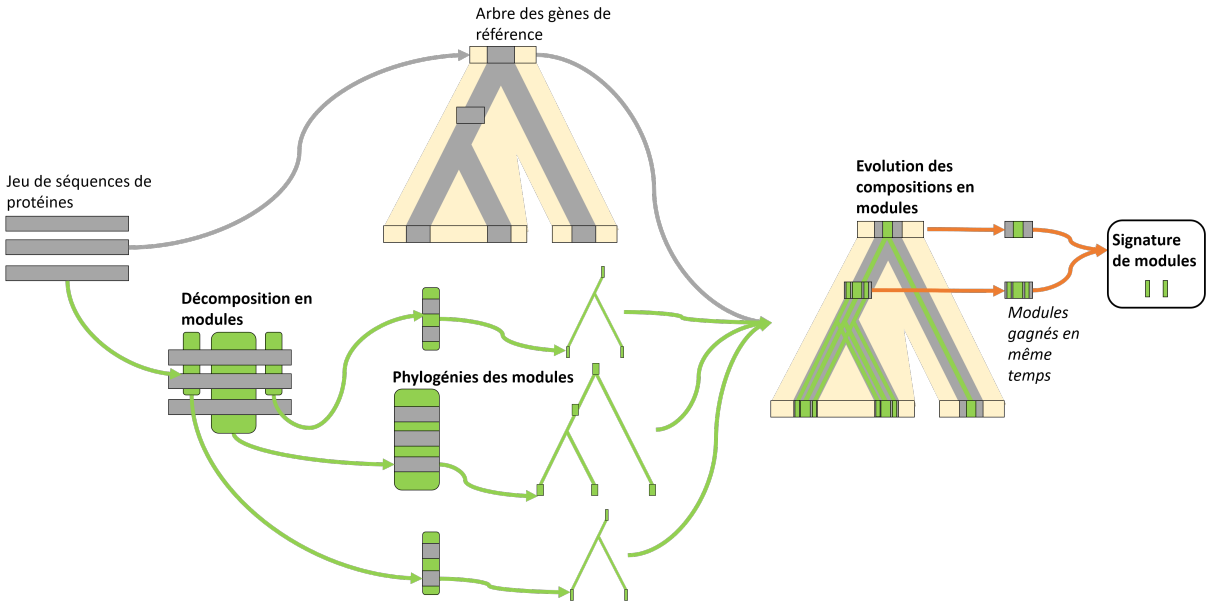

FIGURE 6.1 – **Inférer l'évolution des compositions en modules.**
Un jeu de séquences de protéines homologues, dont la phylogénie est connue, est segmenté en modules conservés. La phylogénie de chaque module est inférée avant d'être réconciliée avec l'arbre des gènes et l'arbre des espèces. La composition ancestrale en module de chaque gène est alors inférée, ce qui permet d'analyser l'évolution des compositions en modules, notamment les gains de modules au cours de l'évolution des gènes.

6.1 Description d'un module de conservation

Nous définissons la notion de *module*, comme une sous-unité des séquences, définie sur la seule base de la conservation forte d'une région des séquences, et dont il est possible d'inférer l'histoire évolutive des segments le composant. Un module est une région conservée par des séquences de protéines orthologues et/ou paralogues, et représente la région d'une séquence fortement conservée qui aurait évoluée plus lentement que le reste de la séquence de la protéine, pour préserver une fonction avantageuse de l'organisme. Un module est un ensemble de segments conservés, d'un sous-groupe de séquences. Nous proposons d'utiliser le programme `paloma-2` [Cos22] (présenté en Section 3.1.3) pour identifier *sans a priori* ces modules dans notre jeu de séquences de références. La difficulté est que notre jeu de séquences de références (214 séquences, 9 espèces) contient des paralogues et des orthologues.

6.1.1 Identification des modules

Dans cette section, nous allons définir la notion de module sur la base d'un PLMA (*Partial Local Multiple Alignement*) et de son découpage en blocs. Dans le but de détecter des régions très conservées, nous utilisons le programme `paloma-2` (détails en Section 3.1.3) avec les paramètres `-m 1 -M 20 -t 10`. Avec ces paramètres, seront alors considérés tous les alignements locaux sans *gaps* d'une longueur minimum de 1 (`-m 1`), d'une longueur maximum de 20 (`-M 20`) et d'un poids fort de similarité significative de 10 ou plus (`-t 10`, [Mor99]). Ces paramètres permettent de segmenter en blocs de conservation très forte, correspondant à des régions des séquences protéiques qui auront évoluées beaucoup plus lentement que le reste de la protéine.

Nous avons ainsi la possibilité d'identifier des régions conservées, qui évoluent plus lentement que le reste de leur séquence, cependant nous voulons également étudier par la suite l'histoire évolutive propre à cette région conservée et issue de la modularité des protéines multidomaines. Notre intention est donc d'inférer la phylogénie de chacune de ces régions conservées, sur la base de la divergence entre ces différents segments, afin d'en capter leur évolution indépendante. Se pose le problème d'utiliser des blocs trop courts ne possédant pas assez de sites pour en estimer la phylogénie.

Les blocs d'un PLMA peuvent être de longueurs très variables, et même possiblement de longueur 1, possédant un unique résidu. Ces blocs très courts sont issus de la détection d'une conservation plus longue, impliquant des segments partagés par un minimum de deux séquences, mais ne sont pas utilisables pour une inférence phylogénétique. Un arbre construit sur une seule position n'aurait pas de sens. Une approche envisagée était de regrouper les blocs directement contigus dans une séquence, de manière à créer des *megablocs*. Cependant, notre but est de trouver une solution qui laisserait le maximum d'indépendance aux blocs, car chacun peut avoir une histoire qui lui est propre. En effet, regrouper des blocs, bien que contigus et courts, a pour effet de diluer le signal phylogénétique issu de chacun d'eux (Figure 6.2).

Il nous semblait ainsi important de ne pas regrouper des blocs, de manière à étudier l'évolution indépendante de chaque bloc. C'est pourquoi nous avons choisi de définir un module comme étant un bloc unique, mais en ajoutant une contrainte de longueur pour filtrer les blocs trop petits. Nous avons choisi de considérer comme modules les blocs d'une longueur minimum de 5 résidus. Ce choix est arbitraire. Il a pour but d'éviter l'inférence de phylogénies pour des modules trop courts. Il est important de préciser que de tels arbres phylogénétiques de modules de longueur 5 verront leur topologie corrigée via

l'utilisation du programme de réconciliation `Treefix`, de manière à réduire au maximum les erreurs stochastiques d'une telle phylogénie. Un module conservé est alors un ensemble de segments de séquences alignées par un bloc de PLMA de longueur minimum 5. Notons que les arbres peu soutenus des modules tendront vers la topologie de l'arbre des gènes.

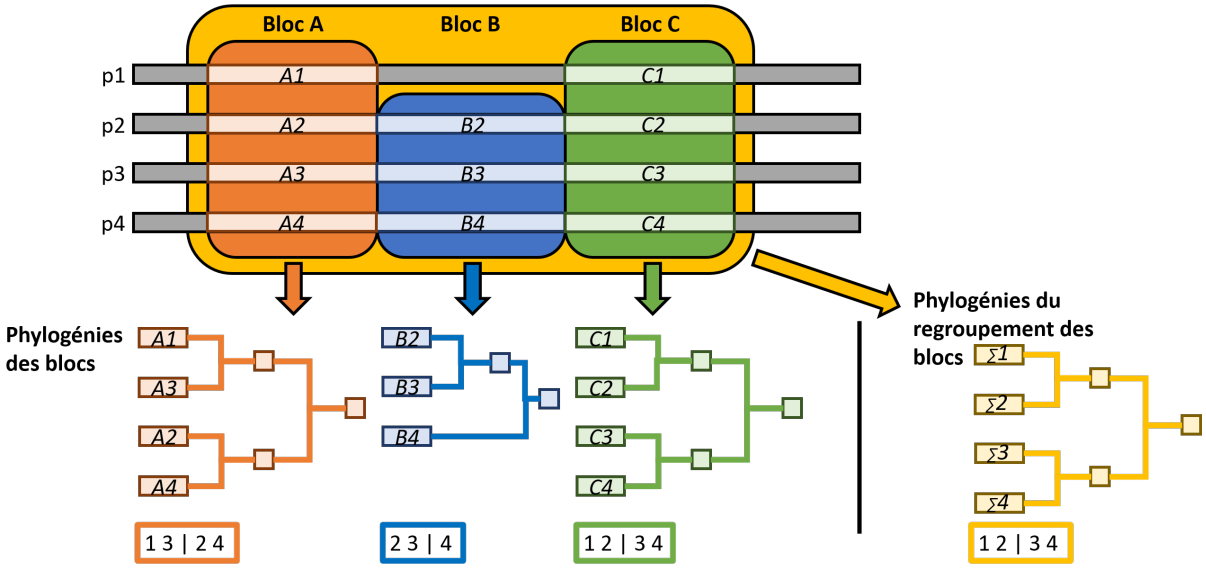

FIGURE 6.2 – **Indépendance des blocs et signaux évolutifs différents.**
Chaque bloc est un ensemble de segments de séquences (e.g., A1, A2, A3, A4 représentent les segments du bloc A). Il est possible d'inférer une phylogénie à partir des segments d'un bloc. Cependant, la phylogénie d'un bloc (orange, bleu, vert) n'est pas nécessairement équivalente à la phylogénie d'un ensemble de blocs dans lequel il serait présent (en jaune), car chaque bloc est susceptible d'avoir une histoire évolutive qui lui est propre et qui est indépendante de l'évolution de la région le contenant.

6.1.2 Visualisation des modules

De manière à faciliter l'interprétation de la localisation d'un module, nous avons mis au point différentes visualisations. Il était important de comparer la localisation des modules avec la localisation des domaines/motifs connus, et nous avons utilisé ceux-ci comme contexte pour positionner les modules sur les séquences. De plus, les domaines constituent une description commune avec les autres études caractérisant les ADAMTS-TSL.

Les domaines et motifs, respectivement Pfam [Mis+21] et Prosite [Sig+02], sont prédits pour chacune de nos séquences. Une protéine est ainsi décrite par sa composition en domaines/motifs (avec leurs coordonnées), et sa composition en modules que nous avons identifiée avec `paloma-2` (également avec les coordonnées de chacun des modules). Nous

avons implémenté une visualisation des modules sur les séquences et des domaines/motifs connus, de manière à savoir si un module est dans un domaine/motif connu, chevauchant deux domaines/motifs consécutifs ou situés à l'extérieur de tout domaine/motif connu. Pour cela, nous avons créé un fichier de description au format de l'outil de visualisation de phylogénies en ligne `Itol` [LB19a] (Figure 6.3). Les domaines sont représentés par des rectangles gris et les motifs « Thrombospondine » par des ronds. Chaque module est représenté par une combinaison couleur/forme différente. Ces visualisations sont générées de manière automatique. De plus, les coordonnées des modules et les segments identifiés peuvent aussi être utilisés pour projeter la localisation d'un module sur une structure 3D (e.g., prédite par `AlphaFold`), dans le but de le localiser dans l'espace.

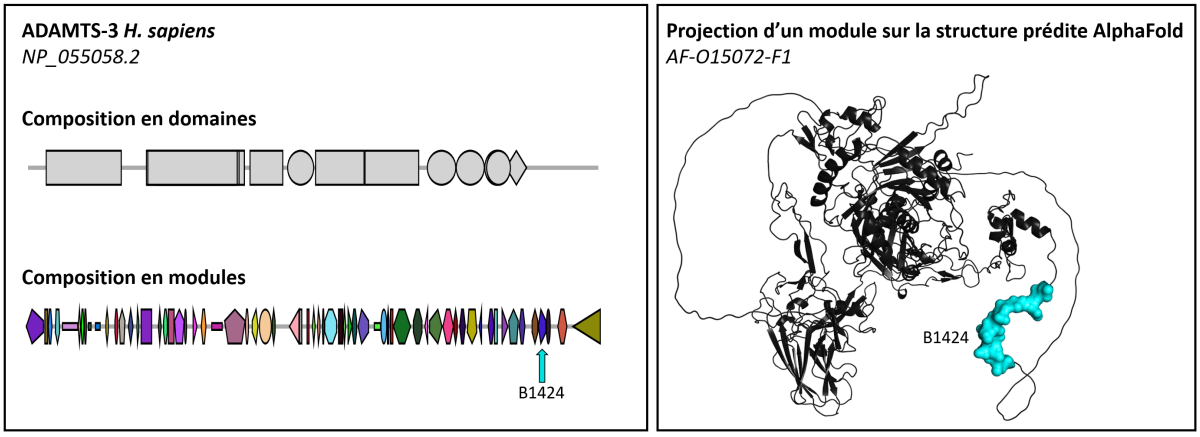

FIGURE 6.3 – **Visualisation de la composition en domaines et en modules de ADAMTS-3 humaine grâce l'outil en ligne `Itol` et projection du module B1424 sur la structure prédite par `AlphaFold`.**

Par ailleurs, décomposer en modules les ADAMTS-TSL nous permet de les segmenter sans *a priori*, et de proposer un nouveau type de décomposition, calculé indépendamment des bases de données, des structures et de tout type d'annotations et spécifiques à différents sous-groupes de séquences. Nous pouvons ensuite comparer et visualiser la localisation de ces modules en comparaison avec les localisations des domaines et des motifs connus.

6.2 Décomposition en modules conservés des 214 ADAMTS-TSL

Nous avons défini un module comme les séquences localement alignées par un bloc de PLMA, dont la longueur est supérieure ou égale à 5. L'outil `paloma-2` nous permet d'identifier les blocs très fortement conservés, qui auraient évolués plus lentement que le reste de la séquence (Section 6.1.1) de manière à obtenir la décomposition en modules d'un jeu de séquences de protéines. Le jeu de séquences *dataset-214* contenant 214 séquences représentatives de gènes homologues de 9 espèces a été segmenté en modules au moyen de `paloma-2` muni des paramètres décrits en Section 6.1.1. La décomposition obtenue contient 1059 modules conservés de segments de longueur minimum de 5 résidus dont la présence au sein des 214 séquences est représentée sur la Figure 6.4.

Cette décomposition en 1059 modules nous permet d'associer à chacune des 214 séquences une composition en module (i.e., une liste de modules présents dans sa séquence). La composition en modules des 26 paralogues ADAMTS-TSL humains est représentée en mosaïques sur la Figure 6.5.

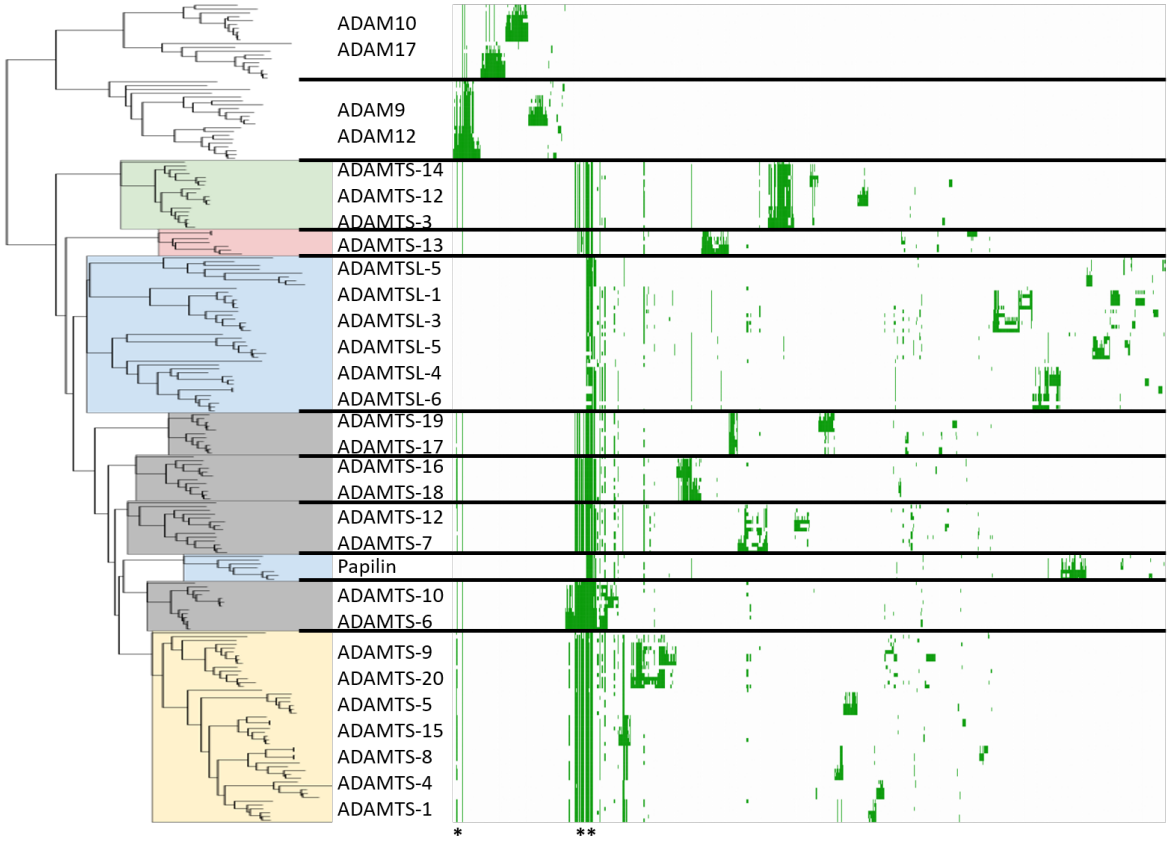

FIGURE 6.4 – **Occurrence des 1059 modules chez les 214 séquences ADAMTS-TSL.** Chaque ligne correspond à une séquence de protéine parmi les 214 séquences ADAMTS-TSL. Chaque colonne correspond à un module (non classé par position sur leurs séquences). La présence d'un module dans une séquence est représentée par un carré vert. Nous observons des modules partagés par les protéines ADAMTS et les ADAM du groupe externe (bandes vertes *) mais absents des ADAMTSL, suggérant des conservations partagées présentes dans le domaine catalytique commun. Un nombre important de modules est partagé par toutes les ADAMTS-TSL (bandes vertes **), alors que la majorité des modules est spécifique à des sous-groupes fonctionnels (colorés sur l'arbre). Figure réalisée avec `Itol`.

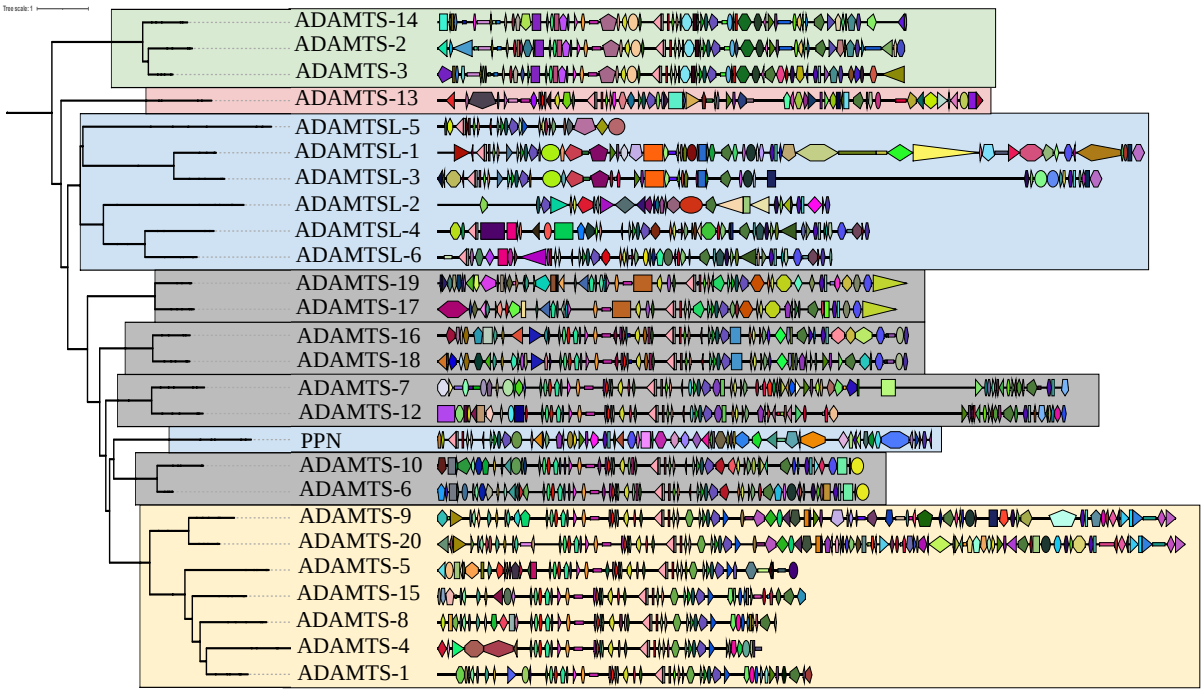

FIGURE 6.5 – **Composition en modules des 26 paralogues ADAMTS-TSL humains.** Chaque module est associé à une combinaison forme/couleur qui lui est spécifique. Issue de la décomposition `paloma-2` en modules des 214 séquences ADAMTS-TSL du *dataset-214*, les segments d'un même module sont représentés par une même forme/couleur. Les protéines sont représentées à l'échelle et les modules sont représentés aux loci correspondants. Figure réalisée avec `Itol`.

6.3 Inférer l'évolution des modules au cours de l'histoire des gènes

Notre objectif est de constituer une carte des modules présents, acquis et perdus à chaque gène de l'arbre des gènes, tout en considérant l'évolution indépendante de chaque module (qui lui est propre et qui résulte de brassage exonique, voir Section 2.3.1) ainsi que son évolution interdépendante (qui résulte de l'évolution de son gène et de son génome). L'évolution indépendante d'un module est modélisée par son arbre phylogénétique et l'interdépendance par une réconciliation phylogénétique mDGS [LB19c] (le *framework* de réconciliation multidomaine le plus abouti à ce jour, voir Section 2.3.3). Ceci permet d'associer les nœuds des différents arbres sur la base d'évènements de duplication, transfert, perte (modèle DTL, *Duplication, transfer, loss*). Les effets du brassage exonique sont bien plus complexes et peuvent inclure des fusions, des fissions ainsi que d'autres évènements pas encore décrits et particulièrement complexes à modéliser dans un modèle de réconciliation [LB19b ; Oak17 ; WRK12]. Ces évènements vont alors être modélisés dans un modèle DTL (comme la réconciliation mDGS) par une grande quantité de transferts, de duplications et de successions de pertes/acquisitions, de manière à réconcilier les arbres sur la base des évènements considérés par le modèle. Nous avons ici choisi d'utiliser une telle réconciliation uniquement dans le but d'associer les nœuds des arbres, et d'interpréter les évènements et *mapping* proposés par la réconciliation mDGS comme des évènements discrets qui permettent de déduire la présence/absence de modules aux gènes ancestraux, que nous modéliserons sous la forme d'une carte de modules présents. Nous proposons alors un concept de carte de l'évolution des modules (Figure 6.6) qui décrit les modules spécifiques pour tous les sous-groupes (bipartitions de l'arbre des gènes) d'ADAMTS-TSL, qui corrige les effets dus au seuil de conservation choisis pour `paloma-2` et qui explique la modularité due au brassage exonique en raisonnant uniquement par présence/absence de modules au sein d'un gène ancestral.

6.3.1 Considérer l'évolution des segments d'un module

Un module est composé de segments de séquences alignés et fortement similaires, ayant divergé de manière indépendante et interdépendante de l'évolution du gène et des espèces. La divergence entre segments d'un même module nous permet d'inférer sa propre histoire évolutive (Figure 6.7).

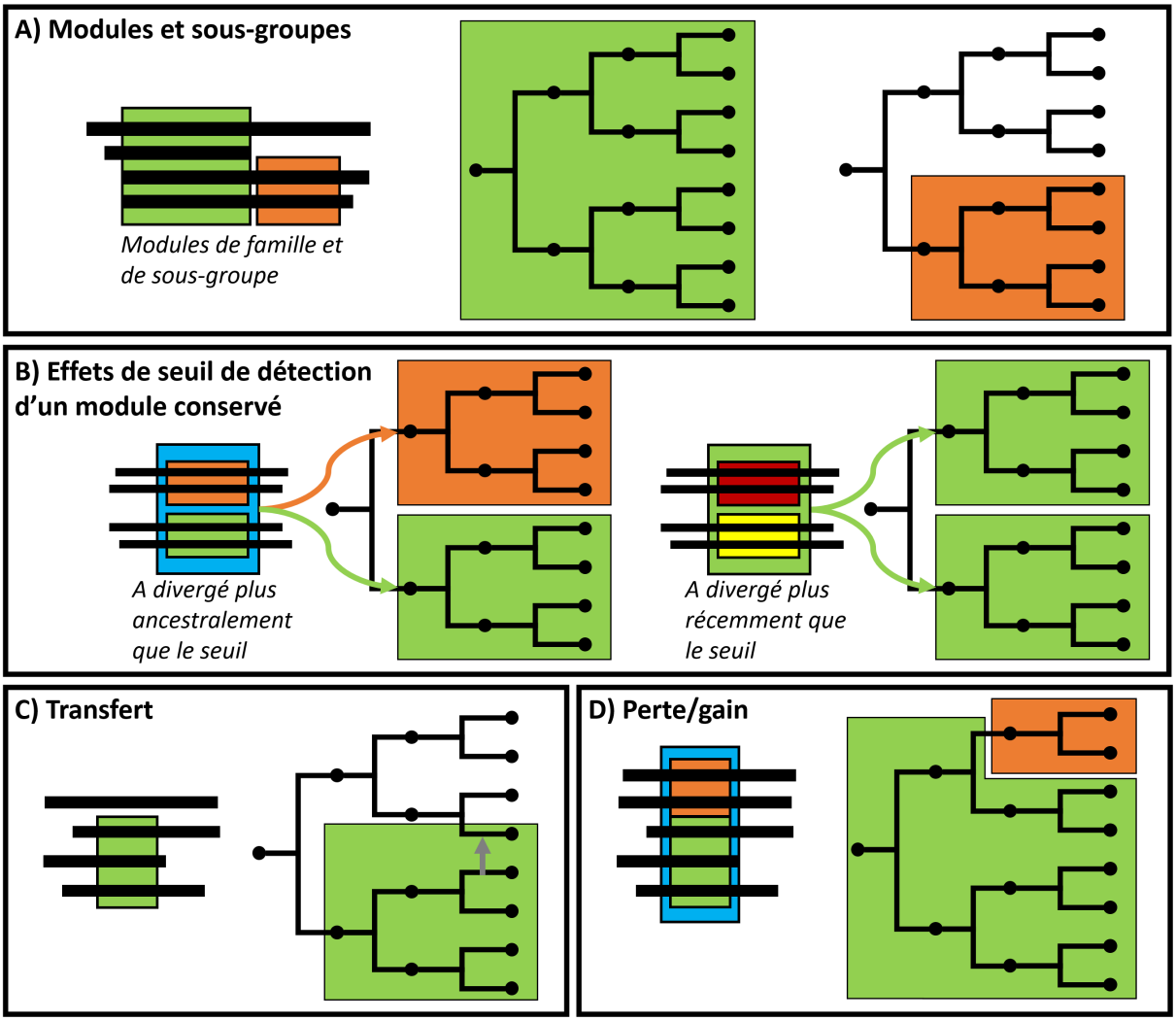

FIGURE 6.6 – **Carte de présence des modules le long de l'évolution des gènes.**
A) Un module peut-être définit comme spécifique à différents niveaux de sous-groupes, et gagné à l'ancêtre correspondant, ici le module vert est spécifique à toutes les séquences et le module orange à seulement un sous-groupe de séquences. **B)** Effets de seuil de détection d'un module conservé. Si un segment ancestral (en bleu) a divergé plus ancestralement que ce que le seuil détecte, deux modules différents décrivent un même locus. Si au contraire un module a divergé plus récemment que ce que le seuil détecte, un unique module décrit deux segments divergents (en jaune et en rouge) qui, par la considération de l'évolution propre du module les regroupant (en vert), permet de capter cette divergence : ici, le module est gagné de manière distincte à l'ancêtre de deux sous-groupes. **C)** Un transfert de module modélise un module spécifique à un sous-groupe, mais également retrouvé dans une séquence lointaine dans l'arbre. **D)** La spécialisation d'un locus donné (en bleu) au sein d'un sous-groupe se modélise par le gain d'un module ubiquitaire à la majorité de la famille (en vert), qui est ensuite perdu en même temps qu'est gagné un module spécifique à un sous-groupe (en orange).

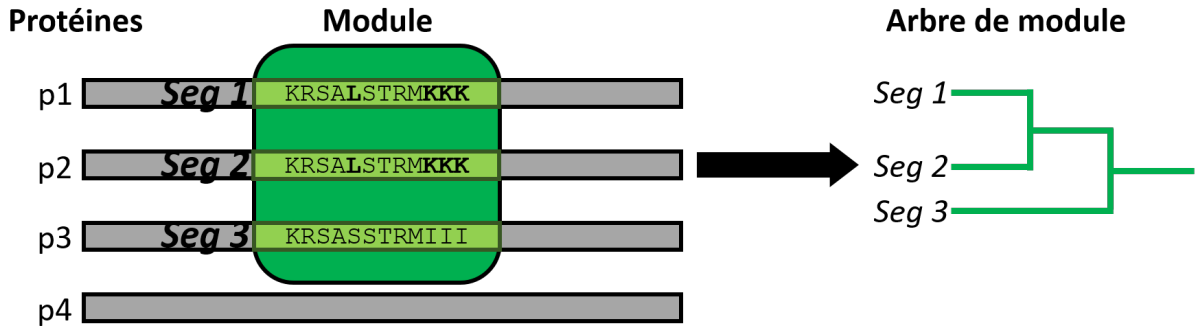

FIGURE 6.7 – **Divergence des segments d'un module.**
Le module contient trois segments (Seg1, Seg2 et Seg3) de trois séquences de protéines (p1, p2, et p3 respectivement), et correspond à une conservation partagée par les protéines p1, p2 et p3, mais pas par p4. Les segments Seg1 et Seg2 ont des séquences plus similaires entre elles que la séquence du segment Seg3 (en gras), ce qui peut se représenter par l'arbre des segments du module (arbre de module).

Des segments de séquences peuvent évoluer indépendamment de leurs séquences (voir Section 2.3.1). C'est pourquoi, nous voulons considérer la possibilité que l'histoire évolutive des segments d'un module et des gènes dont ils proviennent, puissent différer. Considérer le signal évolutif contenu dans les séquences de chacun des segments d'un module permet de tenir compte de la possibilité de tels scénarios, cohérents avec les hypothèses d'évolution indépendante des segments des protéines multidomaines (i.e., brassage exonique).

Dans le but de considérer l'interdépendance entre l'évolution des segments d'un module, l'évolution des gènes et l'évolution des espèces, nous avons choisi d'effectuer une réconciliation Module-Gène-Espèce afin d'associer les nœuds des différents types d'arbres (modules, gènes, espèces). Pour ceci, nous avons appliqué un protocole similaire à ceux des précédentes études de réconciliation Domaines-Gènes-Espèces qui visaient à considérer la divergence au sein des domaines [Sto+15 ; LB19b]. Nos modules remplaceront ici les domaines. Le principe est d'inférer l'arbre phylogénétique d'un module, d'en corriger les erreurs stochastiques, avant de le réconcilier (voir Section 2.3.3) avec l'arbre des gènes et l'arbre des espèces dont le module provient. L'arbre de référence des 214 gènes, ainsi que l'arbre des 9 espèces dont ces séquences proviennent, ont été inférés dans le chapitre précédent (Chapitre 5). Une telle méthodologie permet de considérer l'information propre à chacun des trois niveaux d'évolution (espèce, gène, module), de manière indépendante, au sein d'un unique scénario d'évolution, de manière à considérer l'interdépendance de ces évolutions.

6.3.1.1 Inférence d'arbres phylogénétiques de modules

Nous présentons dans ce paragraphe et sur la Figure 6.8 notre méthodologie d'inférence des arbres phylogénétiques des modules. L'hypothèse considérée ici est la suivante : la séquence d'un gène évolue avec le génome dont il provient (évolution des espèces), mais évolue également au sein du génome. De la même manière qu'un segment génomique (e.g., module) évolue avec le gène dont il provient (interdépendance de leurs évolutions), mais évolue également indépendamment au sein du gène. Cette indépendance des trois niveaux d'évolution (espèce, gène, module) se traduit par de possibles incongruences entre les arbres des différents niveaux d'évolution.

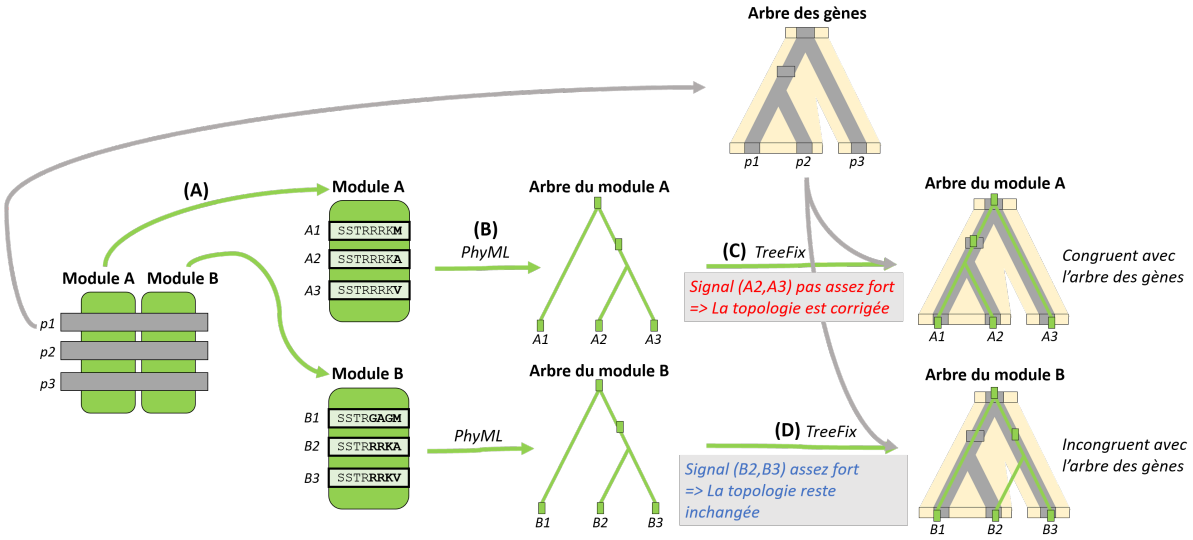

FIGURE 6.8 – **Inférence d'arbres phylogénétiques de modules.**
(A) Les séquences p1, p2 et p3 (dont l'arbre des gènes est connu) possèdent toutes deux modules conservés : le module A et le module B. Le module A est décrit par les trois segments A1, A2 et A3 issus des séquences p1, p2 et p3 respectivement. (B) Inférer la phylogénie des segments A1, A2 et A3 permet de reconstruire l'histoire évolutive propre aux segments du module A, modélisé par l'arbre du module A. (C) et (D) Les arbres de modules sont ensuite corrigés avec l'arbre des gènes afin d'éliminer les incongruences non significatives entre un arbre de module et l'arbre des gènes.

Dans le cas d'un module, les segments sont alignés par construction. Nous n'avons pas besoin de construire un alignement multiple des segments du module (Figure 6.8 (A)), il nous suffit d'inférer une phylogénie (ici avec `PhyML` [Gui+10], Figure 6.8 (B)) afin d'obtenir un arbre phylogénétique du module qui représentera ainsi l'histoire évolutive de ses segments.

Les segments étant souvent de longueurs assez réduites, un arbre de module est ainsi

très fortement biaisé par les erreurs stochastiques. De manière à réduire ces erreurs, nous avons utilisé `TreeFix` [Wu+13] pour corriger l'arbre de module en fonction de l'arbre de référence des protéines dont les segments sont issus. En effet, bien qu'un module puisse posséder une histoire évolutive différente de celle du gène dont il provient, il est plus parcimonieux de considérer que son histoire est celle de son gène. C'est dans ce cadre-là que corriger la topologie de l'arbre de module nous permet de considérer uniquement les divergences des segments suffisamment soutenues statistiquement pour ne pas être issues d'erreurs stochastiques. Dans le cas où le signal de phylogénétique serait trop faible, `TreeFix` optera pour une topologie congruente avec celle de l'arbre des gènes (Figure 6.8 (C)). Il faut noter que la topologie obtenue n'est pas pour autant totalement congruente avec celle de l'arbre des gènes. Des incongruences entre l'arbre du module et l'arbre des gènes, soutenues par des divergences significatives des segments du module, vont subsister (Figure 6.8 (D)), suggérant une évolution indépendante du module. Une incongruence entre l'arbre du module et l'arbre des gènes suggère ainsi un événement évolutif propre au module et indépendant de l'histoire du gène (i.e., brassage exonique). Finalement, utiliser `TreeFix` sur un arbre de module permet d'en corriger la topologie à partir de l'arbre des gènes, dans le but de conserver uniquement les incongruences significatives.

En appliquant la méthodologie présentée ici à chacun des 1059 modules, nous obtenons 1059 arbres phylogénétiques des modules (un arbre par module). Chacun des 1059 arbres phylogénétiques de modules, représente l'évolution indépendante des segments du module et sera par la suite réconcilié avec l'arbre des gènes et celui des espèces (réconciliation Module-Gène-Espèce), dans le but d'associer les nœuds des évolutions qui sont également interdépendantes (i.e., les évolutions des 1059 modules, l'évolution des 214 gènes et l'évolution des 9 espèces).

6.3.1.2 Réconciliation Module-Gène-Espèce

À ce stade, nous avons inféré trois types d'arbres phylogénétiques différents : 1) un arbre de nos 9 espèces, 2) un arbre de référence de nos 214 gènes ADAMTS-TSL ainsi que 3) 1059 arbres de module. Ces trois types d'arbres représentent l'information de trois différents niveaux d'évolution, chacun partiellement indépendant et interdépendant d'un autre. En effet, les gènes évoluent au sein des espèces, et les modules évoluent au sein des gènes, cependant les arbres correspondants ne sont pas nécessairement congruents, témoignage d'événements évolutifs propres à chacun de ces niveaux (détaillé en Section 2.3.1). Notre but est de réconcilier les phylogénies de niveaux différents, afin d'associer les

nœuds des différents arbres sur la base d'un scénario d'évolution cohérent des modules, des gènes et des espèces. Pour ceci, nous utilisons le programme de réconciliation DGS (*Domain-Gene-Species*, détaillé en Section 2.3.3) `SEADOG-MD` [LB19c] afin d'effectuer une réconciliation des phylogénies des modules, des gènes et des espèces (Figure 6.9).

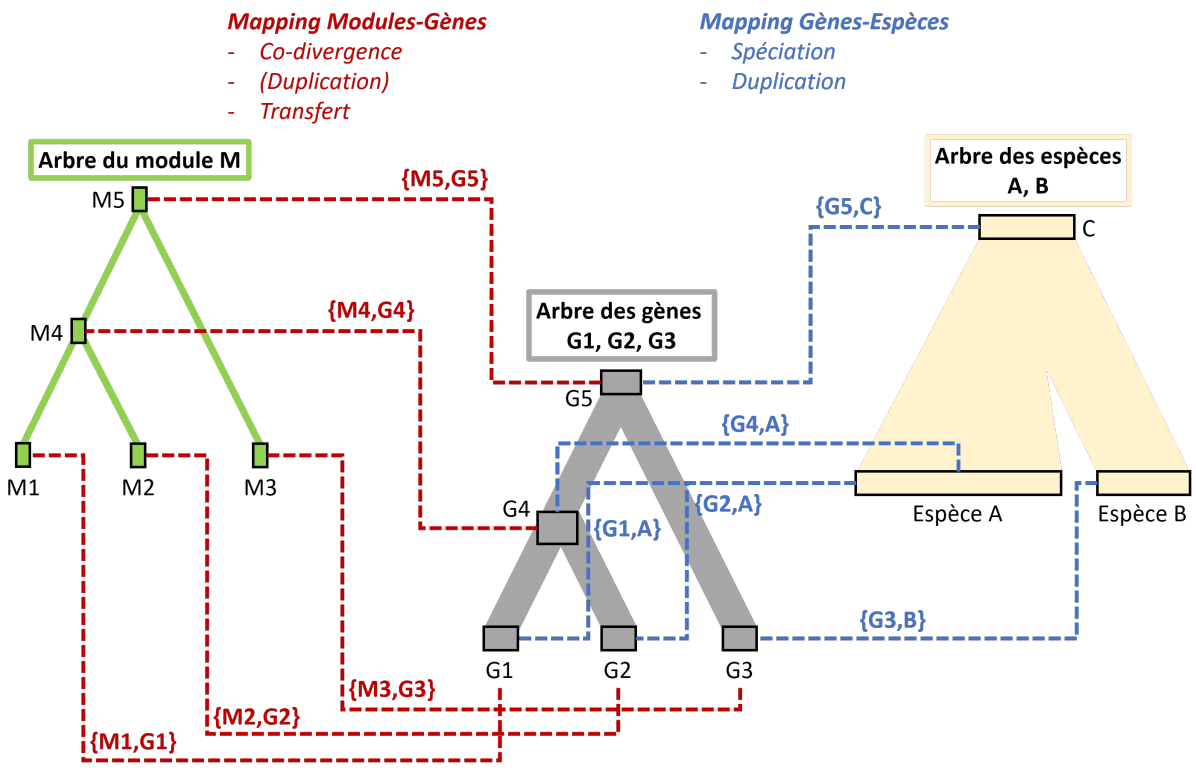

FIGURE 6.9 – **Réconciliation Module-Gène-Espèce.**
Réconcilier un arbre de module, un arbre de gènes et un arbre des espèces avec la méthode mDGS de `SEADOG-MD` [LB19c], permet d'associer chaque nœud de l'arbre de module (segment actuel ou ancestral du module) à un nœud de l'arbre des gènes (gène actuel ou ancestral), tout en associant de manière conjointe chaque nœud de l'arbre des gènes à un nœud de l'arbre des espèces. Les différentes associations entre les nœuds (*mapping*, noté de la manière suivante : `{nœud_1,nœud_2}`) de ces différents arbres sont expliqués par différents types d'événements évolutifs. Sur la figure, G1 et G2 sont issus d'une duplication paralogue dans l'espèce A. L'arbre des modules suit l'arbre des gènes sans événement propre.

La réconciliation multiModule-Gène-Espèce avec le programme `SEADOG-MD` prend en entrée l'arbre des espèces, l'arbre des gènes et les arbres des modules, et nous permet d'obtenir les *mapping Gène-Espèce* ainsi que les *mapping Module-Gène*, sans modifier aucune des topologies. Un *mapping* associe un nœud d'un arbre à un nœud d'un autre arbre. Dans un premier temps, le *mapping* le plus simple est le *mapping de feuille* qui va associer les feuilles des arbres suivant leur appartenance (e.g., un module avec le gène

dont il est issu). Grâce à ces *mapping de feuille*, il est ensuite possible de comparer les bipartitions des différents arbres, et ainsi d'effectuer des *mapping* entre les nœuds internes. Ces *mapping* des nœuds internes permettent d'expliquer la congruence (évolution interdépendante) ou l'incongruence (évolution indépendante) de leurs bipartitions par un événement évolutif. Un *mapping Gène-Espèce* associe un nœud de l'arbre des espèces à un nœud de l'arbre des gènes (en bleu sur la Figure 6.9). Les événements qu'il peut représenter sont soit une *spéciation*, soit une *duplication de gènes*. D'une manière similaire, un *mapping Module-Gène* associe un nœud d'un arbre de module à un nœud de l'arbre des gènes (en rouge sur la Figure 6.9), les événements considérés pouvant être : une *co-divergence*, une *duplication de module*, un *transfert de module*.

Les topologies des arbres de modules pouvant être incongruentes avec celle de l'arbre des gènes, la réconciliation Module-Gène-Espèce nous permet d'associer l'information de ces différents arbres et d'interpréter les incongruences comme des événements évolutifs propres (duplication, perte, transfert). Nous avons alors interprété les *mapping* produits par la réconciliation Module-Gène-Espèce avec `SEADOG-MD`, en particulier les *mapping Module-Gène* afin d'inférer la composition en modules de chacun des gènes ancestraux de notre arbre de référence ADAMTS-TSL et de ne raisonner qu'en termes de présence/absence de modules, et non plus en évènements. Inférer les compositions en modules pour tous les gènes ancestraux revient à inférer l'évolution de la composition en modules ADAMTS-TSL. Le passage des *mapping* de la réconciliation, aux compositions en modules n'est pas trivial et nécessite de traiter l'information de ces *mapping* et d'en interpréter les évènements et les informations discrètes du modèle.

6.3.2 Interpréter la réconciliation en présence/absence de modules aux gènes ancestraux

Notre objectif ici est d'interpréter les *mapping*, l'association des nœuds qu'ils proposent, le sens discret des différents évènements associés et la topologie des arbres, afin de constituer une carte de la présence des modules au sein de l'évolution des gènes. Notre utilisation de la réconciliation Module-Gène-Espèce se limite ainsi à l'interprétation des *mapping*, qui expliquent les évolutions indépendantes et interdépendantes par des associations de nœuds et d'évènements, dans le but de reconstruire les compositions ancestrales en modules des gènes ancestraux (i.e., présence/absence de modules à chaque gène ancestral).

Les *mapping Module-Gène* produits par la réconciliation Module-Gène-Espèce avec `SEADOG-MD` associent chaque nœud d'un arbre de module (segment ancestral du module), à un nœud de l'arbre des gènes (gène ancestral), mais n'indique pas explicitement à quels nœuds ancestraux de l'arbre des gènes le module est présent. Nous avons ici inféré la présence du module aux gènes ancestraux à partir de ces *mapping Module-Gène*, en considérant également la topologie de l'arbre du module et la topologie de l'arbre des gènes, associées par un *mapping Module-Gène*. Nous posons l'hypothèse suivante : si un nœud de l'arbre des gènes (un gène actuel ou ancestral) est associé à un nœud de l'arbre du module (un segment actuel ou ancestral du module), alors ce module est jugé présent au sein de ce gène.

Nous allons alors utiliser les résultats de la réconciliation Module-Gène-Espèce afin d'inférer la composition ancestrale en modules de chaque gène (i.e., les modules présents dans un gène ancestral). Pour ceci, nous utilisons l'arbre du module, l'arbre des gènes, et les *mapping Module-Gène* entre les deux, de manière à confronter et à rassembler l'information contenue dans les deux arbres. Un *mapping Module-Gène* associe un nœud d'un arbre de module à un nœud de l'arbre des gènes. Chaque nœud d'un arbre de module possède un *mapping Module-Gène* l'associant à un nœud de l'arbre des gènes. Les *mapping Module-Gène* nous servent alors à associer l'arbre du module à l'arbre des gènes, tout en considérant leurs évolutions propres. Sur la base de cette hypothèse et pour chacun des types de *mapping Module-Gène* possibles (i.e., *feuille*, *co-divergence*, *duplication*, *transfert*), nous avons mis au point une méthode permettant de l'interpréter et d'inférer les gènes ancestraux où un module est présent, considérant toutes les informations discrètes issues des différents *mapping* et des différentes topologies (Figure 6.10).

Pour résumer, la réconciliation associe via des *mapping* les nœuds des arbres des modules aux nœuds de l'arbre des gènes. Nous proposons de considérer les *mapping* et la topologie des arbres pour inférer la carte des présences/absences des modules aux gènes ancestraux, i.e., leur composition ancestrale en modules.

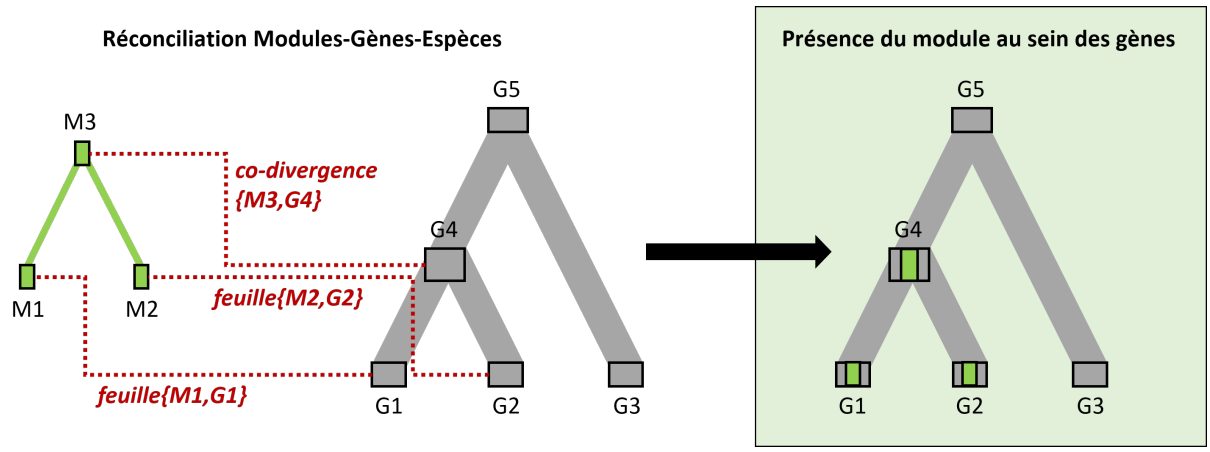

FIGURE 6.10 – **Utilisation de *mapping Module-Gène* pour inférer la présence des modules au sein des gènes.**
Concept de la méthode détaillée en Section 6.3.2, qui utilise l'arbre d'un module M1, l'arbre des gènes et les *mapping Module-Gène* entre les deux, dans le but d'inférer la présence du module M au sein des gènes. Les nœuds G1, G2, G3 correspondent aux gènes actuels, et G4, G5 aux gènes ancestraux. À l'aide des *mapping Module-Gène* représentés en rouge, nous inférons la présence du module M au nœud ancestral G4 et aux gènes G1 et G2.

6.3.2.1 Interprétation de mapping de feuille

Un *mapping de feuille* est un *mapping* trivial entre une feuille d'un arbre de module, et une feuille de l'arbre des gènes. Une feuille de l'arbre de module est un segment de séquence d'une protéine, cette protéine étant une feuille de l'arbre des gènes. Le *mapping* de *feuille* associe donc le segment de module à la protéine dont il provient. Nous considérons ainsi le module comme présent dans le gène associé (exemples avec les *mapping* `{M1,G1}` et `{M2,G2}` sur la Figure 6.10).

6.3.2.2 Interprétation de mapping de co-divergence

Une *co-divergence* d'un module signifie que le module a évolué avec le gène. C'est-à-dire que quand un gène est dupliqué par *spéciation* (i.e., a hérité une copie dans chaque nouvelle espèce), ou par *duplication du gène* (i.e., duplication d'un fragment de génome ou du génome complet, une copie du gène est héritée dans chaque copie), le module est simplement hérité par chaque gène descendant et diverge de manière synchrone avec le gène, leurs évolutions sont interdépendantes.

Dans le cadre d'un *mapping de co-divergence {M3,G3}* (Figure 6.11), le segment de module nommé M3 (un nœud de l'arbre du module M) est présent dans le gène nommé G5 (un nœud de l'arbre des gènes G) associé par le *mapping co-divergence*, mais également au sein des descendants, jusqu'au prochain événement (pouvant également être une *co-divergence*). C'est pourquoi le module est présent chez G5 et chez les gènes possédant les descendants de G5. Pour les retrouver, nous identifions les segments de modules descendants de M3 (qui correspond à G5) dans l'arbre du module : M1 et M2. Le module est ainsi également présent chez leurs gènes associés : G1 et G3 (que nous identifions par leur propre *mapping*, ici des *mapping* de feuilles). La *co-divergence* nous indique que le module a évolué avec les gènes, de G5 à G1, ainsi que de gènes de G5 à G3, c'est pourquoi nous inférons également la présence du module à tous les gènes ancestraux Gi, présents dans la lignée de G5 à G1, ainsi que dans la lignée de G5 à G3 dans l'arbre des gènes. L'utilisation des topologies de l'arbre du module et de l'arbre des gènes nous permet ainsi d'inférer la présence d'un module qui co-diverge, ayant évolué de manière synchrone avec le gène.

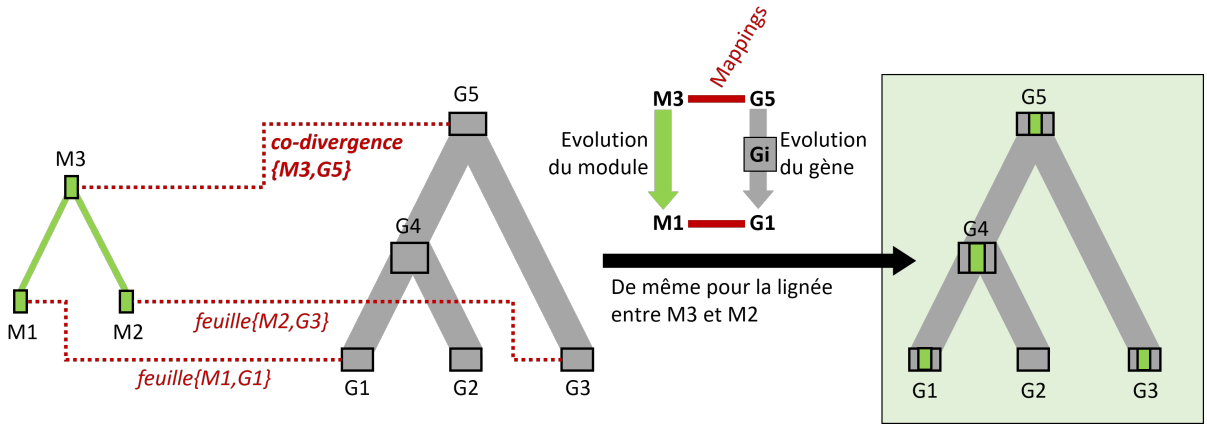

FIGURE 6.11 – **Interprétation du *mapping de co-divergence*.**
La présence du module M en G5 est fournie par SEADOG-MD, du fait que G5 est l'ancêtre commun de G1 et G3 et que M3 est l'ancêtre de M1 et de M2. La présence du module M en Gi est inférée par notre méthode, du fait que Gi est sur la lignée évolutive entre G5 et G1.

6.3.2.3 Interprétation de mapping de duplication de module

Dans le cadre d'un *mapping* de type *duplication de module* (Figure 6.12), un segment d'un module est dupliqué en deux segments de ce module, au sein du même gène. En d'autres termes, un segment ancestral de module dénommé M3, et ses descendants M1 et M2 (également des segments de modules) sont tous associés au même gène G1. Cependant, le module est présent en deux exemplaires au sein de G1. Il est important de noter que compte tenu de la manière dont nous décomposons nos séquences avec `paloma-2`, il est impossible qu'un module soit présent en deux exemplaires au sein d'une même séquence de protéines actuelles (une feuille de l'arbre des gènes). Mais cela n'empêche pas la réconciliation *Module-Gène-Espèce* d'utiliser des événements de *duplication* de modules dans les gènes ancestraux afin d'expliquer des événements propres entre l'arbre d'un module et l'arbre des gènes.

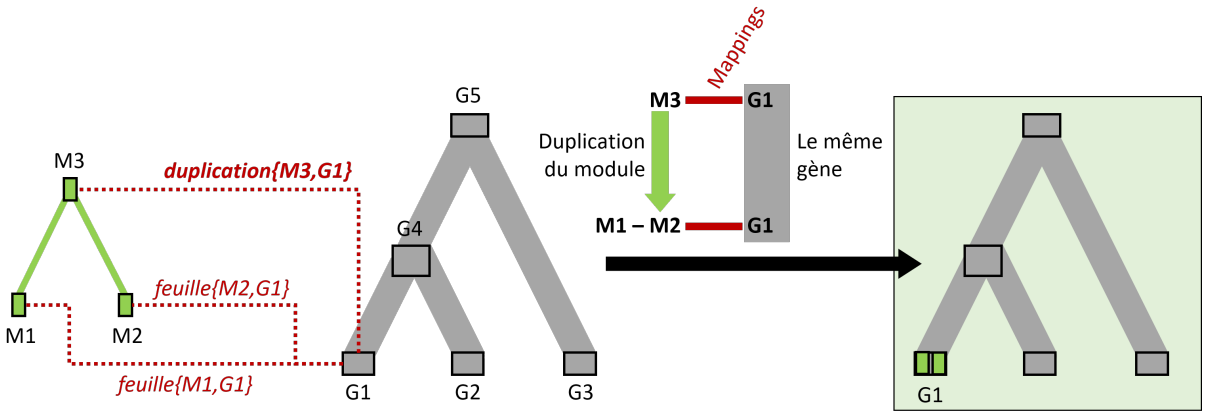

FIGURE 6.12 – **Interprétation du *mapping de duplication*.**
Les segments M1 et M2 du module M sont tous deux présents dans G1, `SEADOG-MD` décrit à leur ancêtre commun M3 un évènement de duplication, également associé au gène G1.

6.3.2.4 Interprétation de mapping de transfert intra-gènes

Le cas le plus complexe est celui du *mapping* correspondant à un *transfert intra-gènes* (Figure 6.13), que l'on nommera simplement *transfert*[1]. Le *transfert* d'un module correspond à la copie d'un module d'un gène au sein d'un autre gène, appelé gène *accepteur*. Le

1. La réconciliation DGS implémentée dans `SEADOG-MD` considère également les transferts inter-gènes, qui correspondent au transfert d'un module d'un arbre de gènes à un autre arbre de gènes. Nous utilisons un seul et unique arbre des gènes, ce qui exclut la possibilité de tels événements.

module apparaît alors chez le gène *accepteur* sans avoir été hérité depuis l'ancêtre. Dans le cas d'un *transfert*, le *mapping* fournit également le nom du gène accepteur, dans notre exemple le *mapping de transfert {M5,G11}, accepteur :G9* décrit le gène G9 comme accepteur. Le nœud de module M5, associé à un événement de *transfert*, comme tout nœud d'un arbre de module, possède deux nœuds de module fils. Le premier M4 correspond à son descendant au sein son gène d'origine (co-divergence), le second M3 correspond au module transféré au gène *accepteur*, distant dans l'arbre des gènes. Grâce à leur *mapping* associés, nous identifions G11, G7 et G6 comme gènes associés aux segments de modules M5, M4 et M3 respectivement. Dans le cas de M4, le module a *co-divergé* de G11 à G7, nous inférons ainsi la présence du module à tous les gènes Gi présents entre G11 et G7. Pour ce qui est du cas de M3, le gène accepteur peut être le gène associé au fils « transfert » ou être un de ses ancêtres s'il y a eu co-divergence après le transfert. Si le gène accepteur est le gène associé au fils « transfert » (ici, si le gène accepteur était G6), nous inférons simplement la présence du module à ce gène. Si ce n'est pas le cas, et que le gène accepteur (G9) est alors un ancêtre du gène associé au fils « transfert » (G6), cela signifie que le module a bien été transféré à G9 et qu'il a *co-divergé* depuis, jusqu'à son prochain événement qui correspond à M3, et donc au gène G6. Nous inférons alors la présence du module à tous les gènes Gi présents entre G9 et G6.

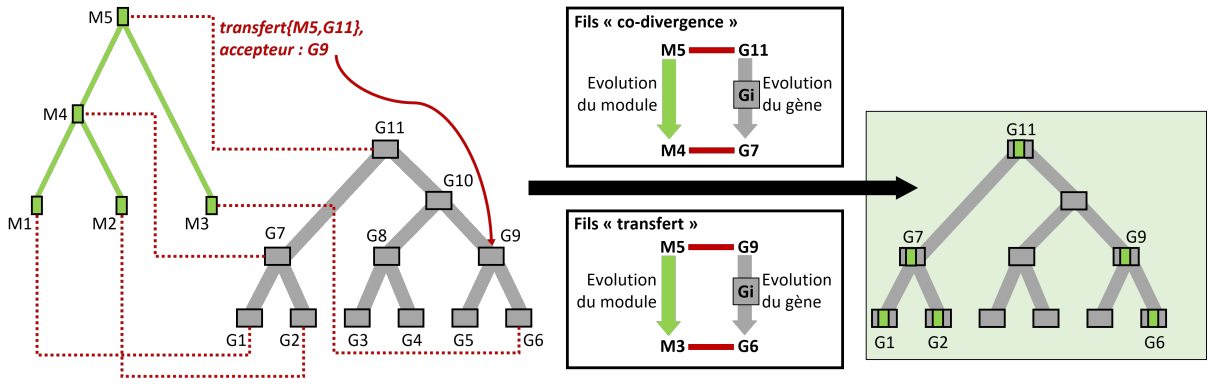

FIGURE 6.13 – **Interprétation du *mapping de transfert*.**
Le segment ancestral de module M5 est transféré au gène accepteur G9. Ses deux segments de modules fils M4 et M3 correspondent respectivement à l'évolution du module au sein de son gène (fils co-divergence) et à l'évolution du module transféré (fils transfert). Le gène accepteur étant un ancêtre du gène associé au fils transfert (G6), le module a également co-divergé entre le gène accepteur du transfert (G9) et le gène associé au segment de module fils descendant du transfert (G6).

6.3.2.5 Implémentation et inférence des compositions ancestrales en modules des ADAMTS-TSL

Nous avons implémenté notre approche d'interprétation de la réconciliation au sein d'un script `Python3`[2] qui nous permet ainsi, sur la base de la réconciliation Module-Gène-Espèce par `SEADOG-MD`, d'inférer la présence de nos 1059 modules au sein des 213 gènes ancestraux de l'arbre de référence ADAMTS-TSL (214 feuilles). Ce qui revient à déterminer la composition ancestrale en modules de chacun des 213 gènes ancestraux (i.e., un gène est associé avec la liste des modules qu'il possède).

2. Fonction `module_gene_inference` du script `integrate_3phylo.py`

6.4 Arbre des gènes et évolution des compositions en modules

Chacun des 427 gènes (213 ancestraux et 214 actuels) de l'arbre de référence ADAMTS-TSL est ainsi décrit par une composition en modules. Nous utilisons la topologie de l'arbre de référence afin d'observer les changements en termes de contenu en modules d'un gène à son descendant (Figure 6.14). Pour ceci, nous comparons les compositions en modules d'un gène et de son ancêtre direct, ce qui nous permet d'identifier trois scénarios possibles : 1) un module est présent chez un gène et chez son ancêtre, 2) un module est présent chez un gène, mais absent chez son ancêtre (*module gagné* chez le gène), et 3) un module est absent chez un gène, mais est présent chez son ancêtre (*module perdu* chez le gène). Notre idée est d'observer uniquement les changements (i.e., l'évolution de la composition en modules) et de définir pour chaque gène la liste de ses *modules gagnés* ainsi que la liste de ses *modules perdus*. Nous définissons ainsi un gène par sa composition en modules, mais également par ses *modules gagnés* et ses *modules perdus* (Figure 6.14).

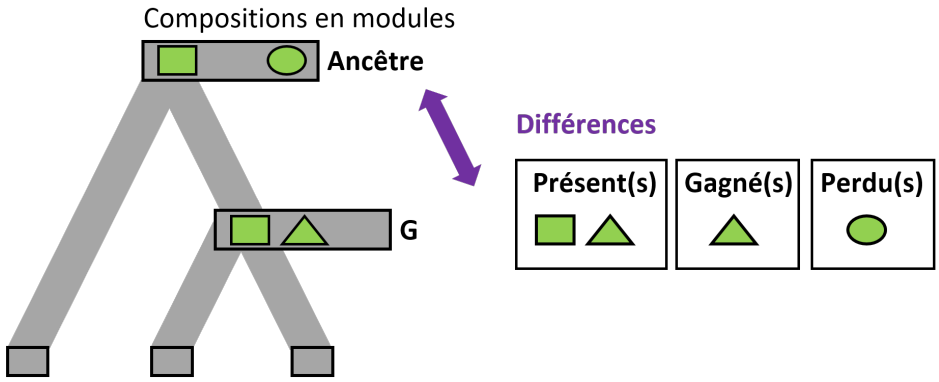

FIGURE 6.14 – **Evolution des compositions en modules.** Une gène ancestral est décrit par la liste de ses modules présents, la liste de ses modules gagnés et la liste de ses modules perdus, compte tenu de la composition en modules de son ancêtre.

6.4.1 Visualisation quantitative des modules gagnés/perdus sur l'arbre de référence ADAMTS-TSL

Dans le but d'analyser rapidement ces gains et pertes, nous avons quantifié le nombre de modules gagnés/perdus pour chaque gène, avant de construire un fichier d'annotation

`Itol` nous permettant de les visualiser sur l'arbre de référence (Figure 6.15). On observe une grande quantité de gains de modules (en vert) au niveau de gènes internes et des gènes représentant les ancêtres pour un ou plusieurs paralogues spécifiques. À l'opposé, on observe un plus grand nombre de pertes de modules (en rouge), tout particulièrement proches des feuilles, et spécifiquement dans certains clades. En effet, les ancêtres récents (nœuds internes proches des feuilles) des gènes ADAMTSL-1, -3, et des gènes ADAMTS-9, -20 présentent un grand nombre de pertes de modules comparé au reste de l'arbre. Ces pertes peuvent s'expliquer par des modules gagnés tôt durant l'évolution (spécifiques à un clade complet) qui seraient ensuite perdus plus tardivement (absents dans un sous-clade). L'absence d'un module dans une séquence (non détecté par `paloma-2`) appartenant à un clade où le module est globalement présent est alors interprété par le modèle comme une perte du module, qui dans certains cas peut être remplacé par un autre module. Ce module de substitution est alors considéré par le modèle comme gagné (représentant une forte divergence de la région, scénario D de la Figure 6.6).

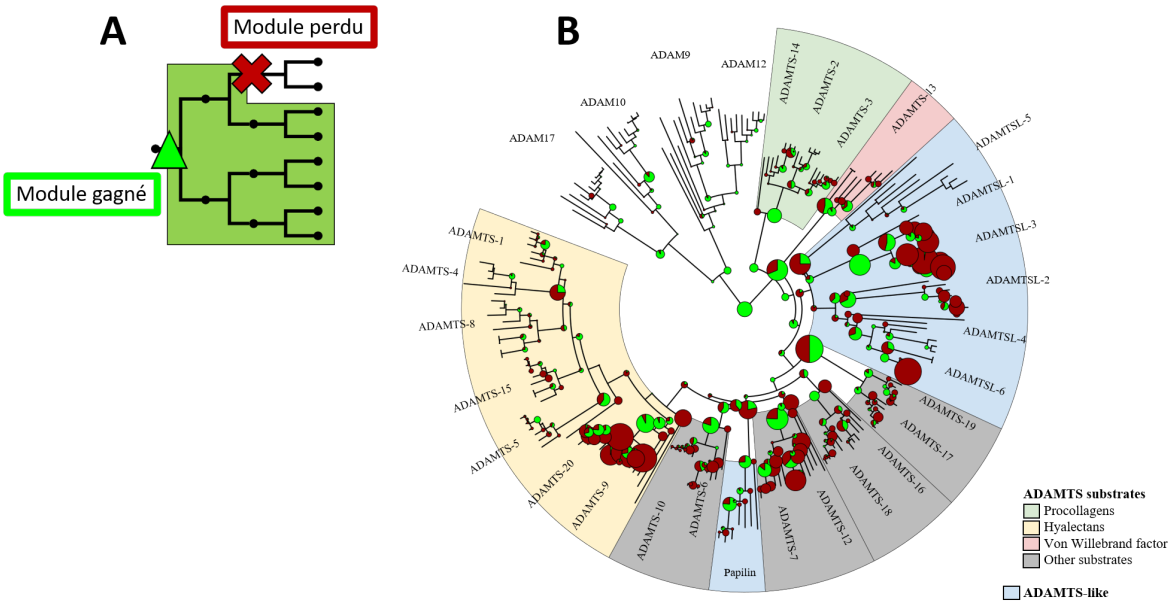

FIGURE 6.15 – **Visualisation quantitative du nombre de modules gagnés/perdus sur l'arbre de référence.**
A) Représentation schématique d'un évènement de gain et d'un évènement de perte d'un module. **B)** Arbre des gènes avec le nombre de modules gagnés représenté en vert et le nombre des modules perdus en rouge. La taille du cercle est proportionnelle au nombre de modules gagnés et perdus.

6.4.2 Visualisation des modules présents, gagnés et perdus à un gène ancestral

Afin de caractériser des ensembles de modules actuels sur la base de leur histoire évolutive, nous avons étudié et visualisé les modules au sein des séquences actuelles les possédant (celles où `paloma-2` les a détectés). Nous voulons nous intéresser plus en détail au contenu de ces listes de modules. Cependant, notre méthode n'infère pas les séquences ancestrales des gènes, ni leur décomposition en modules : nous possédons l'information de la présence d'un module au sein d'un gène ancestral, mais nous ne possédons aucune information sur la séquence du segment de module ancestral correspondant. Ainsi, notre modèle ne nous permet pas de visualiser les modules au sein des gènes ancestraux. Nous avons ainsi mis au point des visualisations `Itol` (Figure 6.16), nous permettant de visualiser, pour un gène ancestral, les modules présents, les modules gagnés et les modules perdus, et ceci, sur les séquences actuelles possédant ces modules.

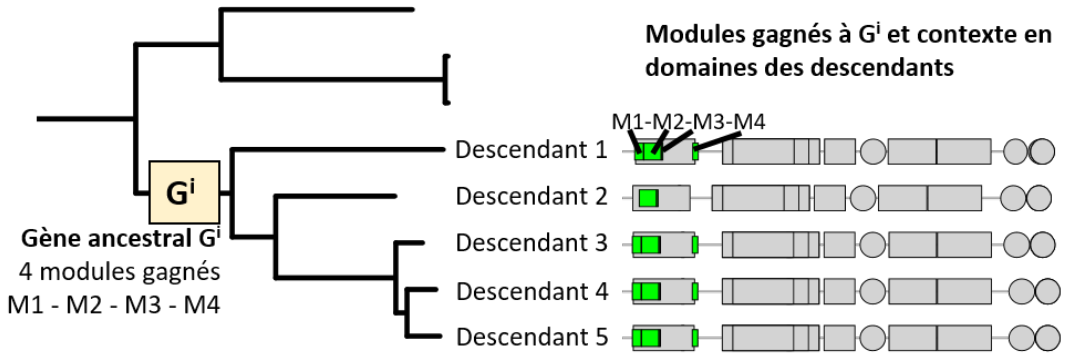

FIGURE 6.16 – **Exemple de visualisation de modules gagnés chez un ancêtre et partagés par les descendants d'un gène.**
Les segments des modules gagnés au gène ancestral G^i sont représentés en rectangles verts sur les séquences des descendants actuels. On remarque ici que les modules M1 et M4 sont perdus chez le descendant 2.

Nous avons réalisé différentes visualisations `Itol`, nous permettant de représenter les modules caractérisant un gène ancestral dans le contexte des domaines. Pour un gène donné, et pour un type de modules (i.e., présents, gagnés, perdus), nous construisons un fichier d'annotation `Itol`. Chacun de ces fichiers représente les segments des modules sous forme de rectangle, brun pour les présents, vert pour les gagnés et rouge pour les perdus. La Figure 6.17 présente un exemple, pour le cas des modules présents et gagnés à

G307, l'ancêtre des orthologues ADAMTS-13. Sur la Figure 6.18 qui présente les modules gagnés à G134, l'ancêtre des papilins, l'un des modules gagnés est également présent dans le clade voisin. Ce qui signifie une divergence particulière de ce module à G134, et donc reflète une spécificité des segments du module chez les descendants de G134, différents des segments présents dans le clade voisin.

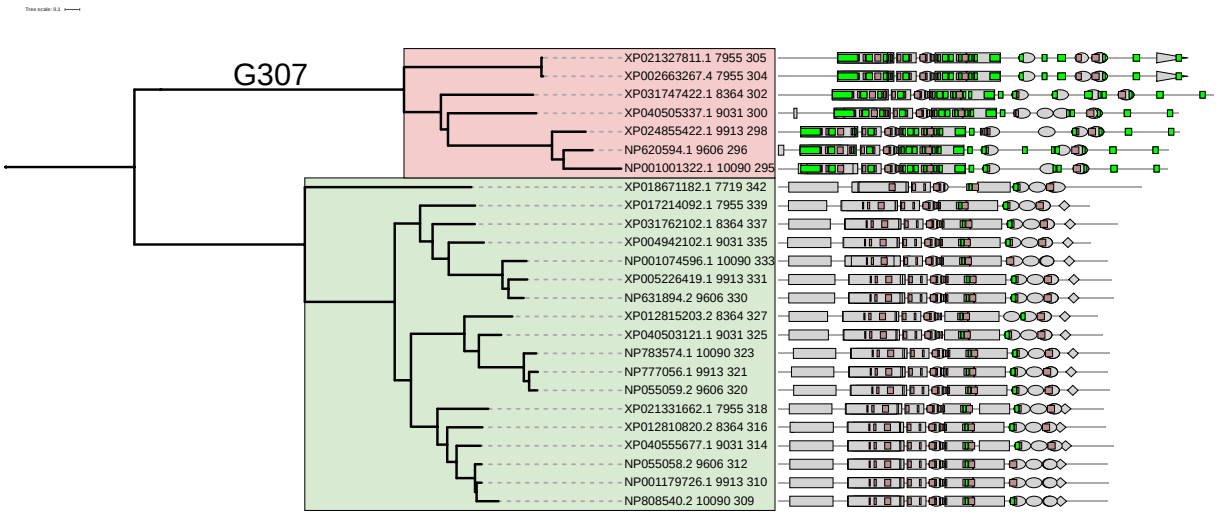

FIGURE 6.17 – **Modules présents et gagnés à G307, l'ancêtre des orthologues ADAMTS-13.**
Seuls les clades des orthologues ADAMTS-13 (en rouge), et des procollagénases (en vert) sont représentés. Les segments des modules présents à G307 (ancêtre des orthologues ADAMTS-13) sont représentés avec des rectangles marron sur la séquence de descendants actuels, les segments des modules gagnés sont représentés par des rectangles verts. Les formes grises représentent les domaines/motifs Pfam/Prosite.

6.4.3 Signature de modules

Nous avons fait le choix de nous focaliser sur les modules gagnés au cours de l'évolution des ADAMTS-TSL (exemple sur la Figure 6.19 avec les modules gagnés au gène ancestral G187). Nous définissons ainsi une *signature de modules* comme l'ensemble des modules gagnés à un gène. L'intérêt d'une signature de modules est d'associer les conservations de séquences apparues à un même moment de l'évolution des gènes. Nous posons ainsi l'hypothèse que les modules composant une signature de modules ont été soumis à une forte pression de sélection depuis un même ancêtre, pouvant être liée à l'apparition d'un phénotype particulier à ce même moment de l'évolution des gènes. Parmi les 213 gènes ancestraux de notre arbre de référence, 186 sont associés avec un gain de module (gain d'un module ou plus), définissant 186 signatures de modules.

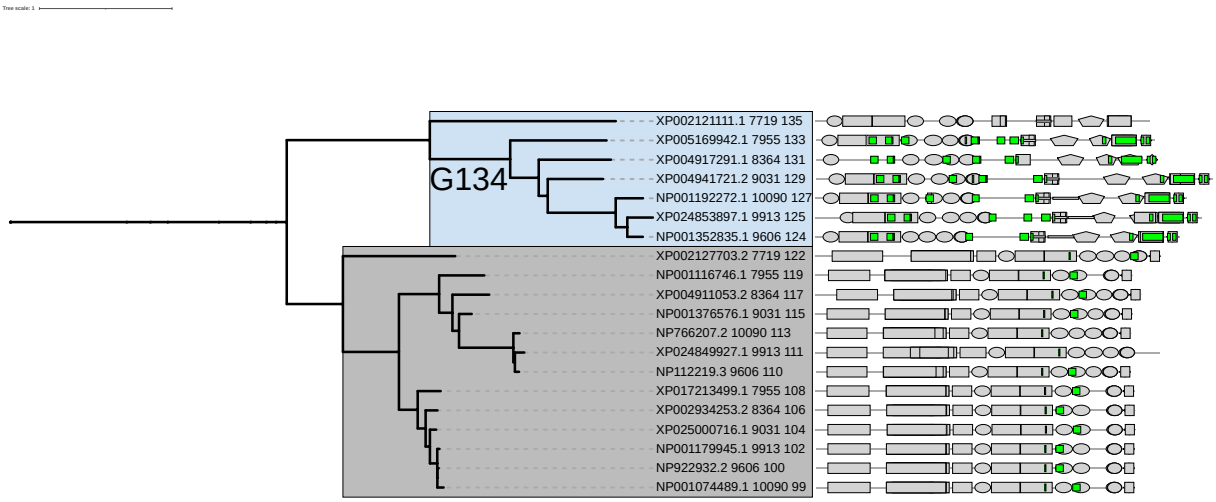

FIGURE 6.18 – **Modules gagnés à G134, l'ancêtre des papilins vertébrés.**
Seuls les clades des orthologues de la papilin (en bleu), et des orthologues ADAMTS-6, -10 (en gris) sont représentés. Les segments des modules gagnés à G134 (ancêtre des papilins vertébrés) sont représentés par des rectangles verts sur la séquence de descendants actuels.

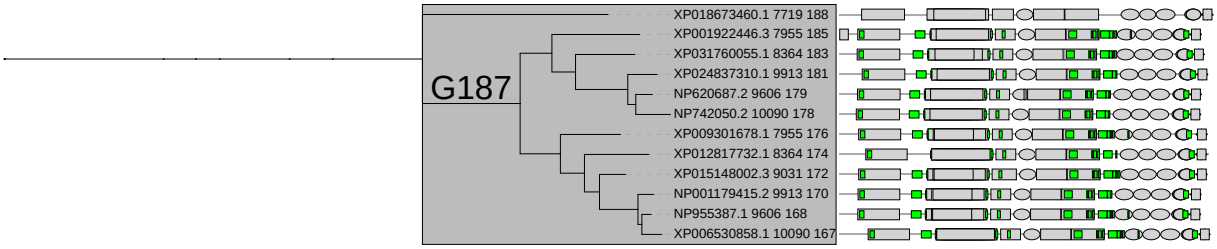

FIGURE 6.19 – **Signature de modules de G187, l'ancêtre des gènes ADAMTS-16, -18 vertébrés.**
Seul le clade des orthologues des gènes ADAMTS-16, -18 (en gris) est représenté. La signature de modules de G187 contient les 15 modules qui y sont gagnés. Les segments de ces modules sont représentés en vert sur les séquences des descendants actuels.

Une signature de modules décrit un signal évolutif partagé par des segments de séquences, c'est-à-dire un ensemble de modules gagnés à un même gène et qui représente un ensemble de conservations qui seraient apparues/auraient divergées de manière synchrone au cours de l'évolution des gènes. Une particularité notable des modules dont l'acquisition est concomitante, est qu'ils décrivent un motif comprenant des modules qui ne sont pas nécessairement contigus, la seule contrainte de regroupement étant qu'ils aient été soumis à une forte pression de sélection depuis un même ancêtre (Figure 6.20).

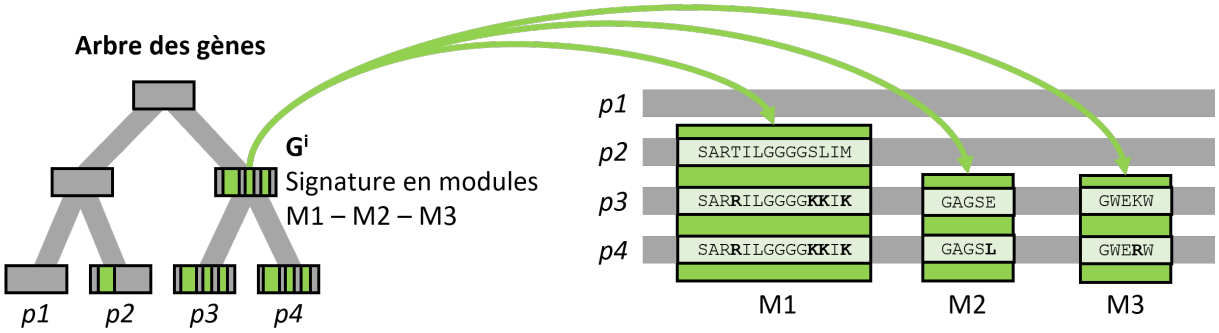

FIGURE 6.20 – **Une signature de modules regroupe des segments soumis à une forte pression de sélection depuis un même ancêtre.**
Le triplet de modules (M1, M2, M3) est gagné au G^i, et constitue sa signature de modules. Contrairement à M2 et M3, le module M1 n'est pas spécifique aux descendants *p3* et *p4* de G^i. Cependant, la divergence au sein des segments de M1 traduit une histoire évolutive de M1 où ce dernier serait apparu deux fois : à G^i et à *p2*. Le gain de M1 à G^i représente ici le gain d'une version divergée de M1 qui est spécifique aux descendants *p3* et *p4*.

6.4.3.1 Visualisation d'une signature de modules

Sur le principe de la visualisation des modules présentée en Section 6.1.2, nous avons développé deux méthodes pour visualiser une signature de modules (Figure 6.21) : 1) les modules peuvent être localisés au sein de leur contexte en domaines, et 2) les modules peuvent être projetés sur une structure tridimensionnelle (e.g., une structure prédite par `AlphaFold`).

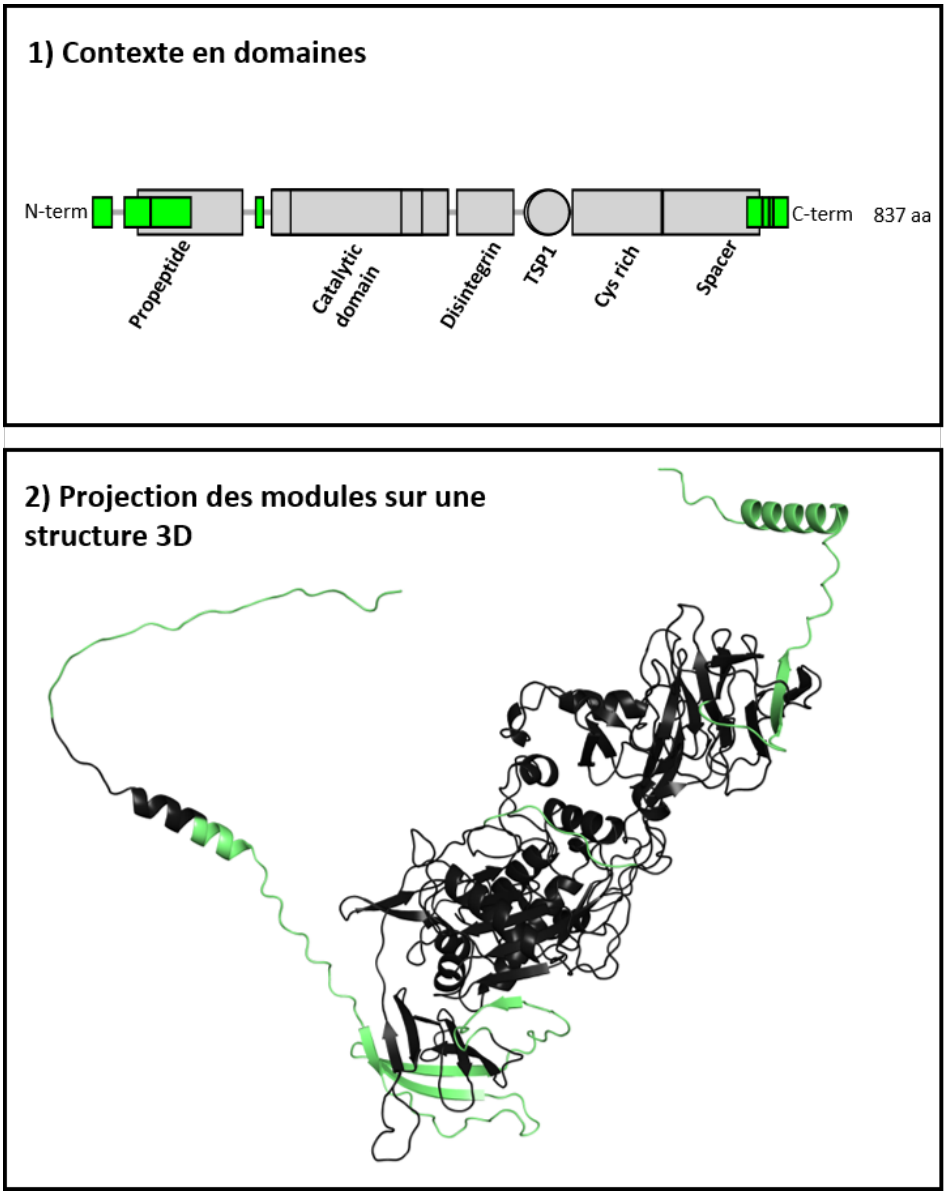

FIGURE 6.21 – **Visualiser une signature de modules.**
Les modules gagnés à un même gène ancestral sont représentés en vert chez une protéine descendante actuelle sur 1) la représentation de sa séquence avec les domaines Pfam, et 2) sur sa structure tridimensionnelle (ici, une prédiction `AlphaFold`).

6.4.3.2 Automate d'une signature de modules

Une signature de modules a pour but de décrire un motif fonctionnel, et comme tout motif fonctionnel, il est intéressant de pouvoir identifier d'autres protéines (non considérées) qui posséderaient également ce même motif. Cependant, une signature de modules représente un enchaînement de modules, discontinu le long des séquences des descendants la possédant. Et bien que ces propriétés (i.e., modulaire, ordonnée, discontinuité, possibilité qu'un module soit absent chez un descendant) rendent le concept des signatures de modules très intéressant, elles complexifient énormément leur utilisation pour requêter de nouvelles séquences. Dans le but de pouvoir requêter avec les signatures de modules, nous avons représenté ces dernières sous forme d'automates (Figure 6.22). L'automate d'une signature de modules représente l'enchaînement ordonné des modules de la signature, leurs segments chez les différents descendants, avec la possibilité de représenter les sauts de modules chez les descendants. Chaque module est composé de différents segments et peut ainsi être représenté par un logo ou un consensus de ces segments. La distance entre deux modules est un intervalle allant de la distance minimale à la distance maximale entre ces deux modules chez les différents descendants. Un automate de signature de modules est alors décrit comme un *protomaton* de la suite `Protomata` [Ker08] (voir Section 3.1.3), ce qui nous permet d'utiliser l'outil `Protomatch` afin de scanner de nouvelles séquences avec un automate de signature de modules.

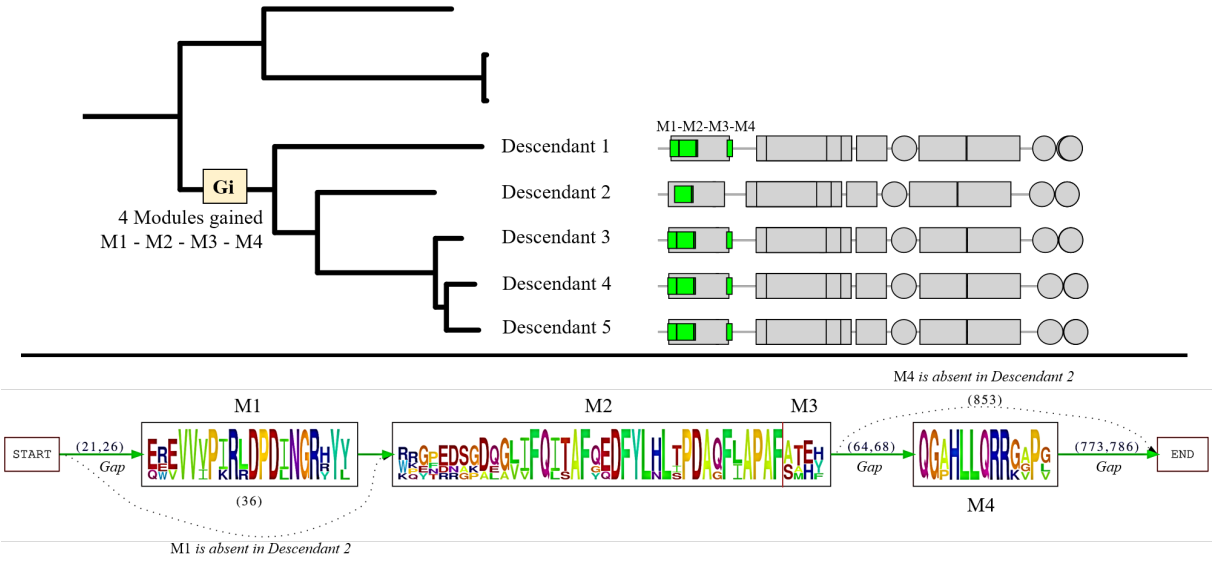

FIGURE 6.22 – **Automate d'une signature de modules.**
La signature de modules du gène ancestral G regroupe les quatre modules M1, M2, M3 et M4 qui sont gagnés à ce gène ancestral. Les modules M1 et M4 sont absents de la séquence du descendant 2. L'automate de la signature de modules (partie inférieure) représente l'enchaînement de ces quatre modules au sein des cinq descendants, et est dirigé de l'extrémité N-term à C-term des descendants. Un module est ici représenté par le logo de ses segments. La distance entre deux modules est représentée par une flèche verte où est indiquée la distance minimale et maximale entre les deux modules chez les descendants. L'absence d'un module chez un descendant est symbolisée par une flèche en pointillé qui saute un module, représentant une transition de l'automate.

Conclusion

Nous proposons une méthodologie novatrice afin d'inférer l'évolution des compositions en modules de gènes ancestraux. Cette méthodologie se base sur une décomposition sans *a priori* des séquences de protéines en modules et considère l'histoire évolutive propre à chaque module. Ces histoires évolutives nous permettent d'inférer l'évolution des compositions en modules des gènes ADAMTS-TSL ancestraux de l'arbre de référence, et ainsi d'étudier la présence, les pertes et les gains de modules ancestraux. Ce sont les gains de modules qui nous intéressent tout particulièrement et nous appelons signature de modules l'ensemble des gains de modules chez un ancêtre donné. Une signature de modules regroupe des segments de séquences, conservés, possiblement non contigus et acquis de manière synchrone au sein de l'évolution des gènes. Appliqué à nos 214 protéines ADAMTS-TSL, notre modèle a permis d'identifier 183 signatures de modules.

Chapitre 7

Évolution des phénotypes des ADAMTS-TSL

Ce chapitre concerne le traitement des phénotypes correspondants aux séquences ADAMTS-TSL. En effet, après avoir identifié les éléments de séquences (modules) spécifiques aux différentes protéines ADAMTS-TSL, mes travaux ont visé à associer des phénotypes aux modules. Pour ce faire, il s'agit d'abord de 1) construire un jeu de données des phénotypes correspondants aux séquences ADAMTS-TSL, puis 2) d'en inférer les histoires évolutives au sein de l'arbre ADAMTS-TSL de référence, dans le but de pouvoir par la suite associer ces phénotypes ancestraux à des modules ancestraux. Les ADAMTS-TSL sont principalement caractérisées par leur capacité à interagir avec d'autres protéines de la matrice extracellulaire. C'est pourquoi les phénotypes que nous avons choisi d'étudier sont les Interactions Protéine-Protéine (PPI). Au cours de ce chapitre (Figure 7.1), il sera question de récupérer un ensemble robuste et le plus complet possible de PPI impliquant les protéines ADAMTS-TSL humaines, de constituer un jeu de données de PPI, puis d'utiliser ces PPI pour annoter nos séquences ADAMTS-TSL avec leurs partenaires d'interactions, de manière à pouvoir finalement inférer l'évolution de ces interactions le long de l'arbre ADAMTS-TSL de référence.

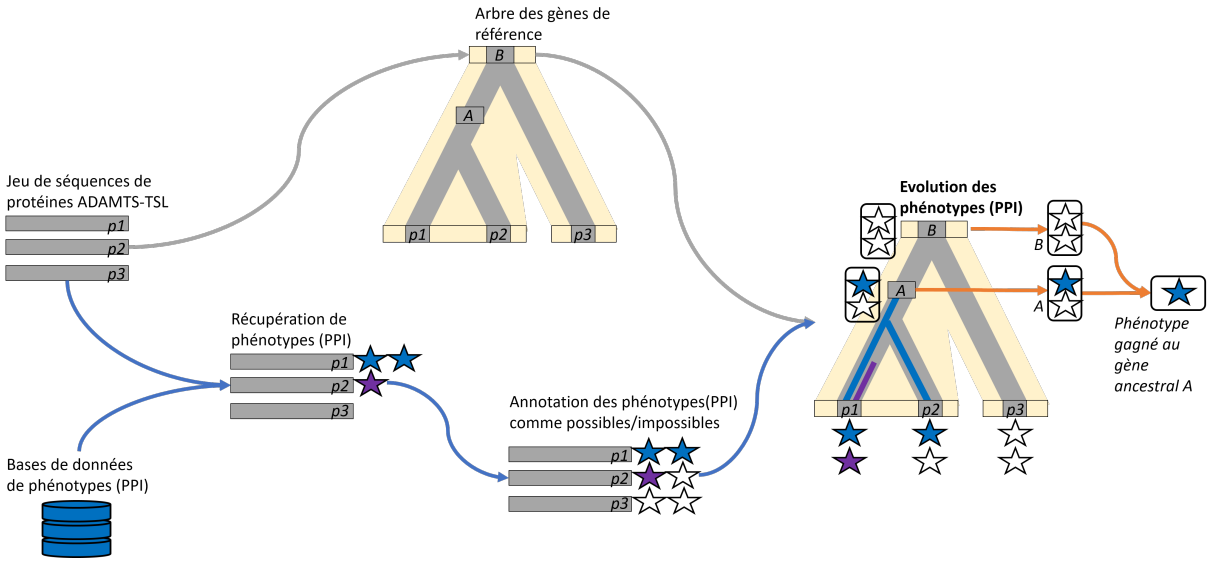

FIGURE 7.1 – **Inférer l'évolution des phénotypes.**
Les phénotypes, ici les PPI, permettent d'annoter les séquences des protéines ADAMTS-TSL (p1, p2, p3) avec pour chacune, sa capacité à interagir (interaction possible) ou non (interaction impossible) avec différents partenaires (les étoiles bleues et violettes représentent des interactions possibles). Il est donc possible d'inférer l'évolution de ces interactions au cours de l'évolution des gènes et d'associer à chaque gène ancestral (A, B) la présence ou non des différentes interactions. Dans notre exemple, l'ancêtre A est connu comme capable d'interagir avec le partenaire représenté par l'étoile bleue, mais pas avec le partenaire représenté par l'étoile violette. Les interactions avec les partenaires bleu et violet sont jugées comme impossibles pour l'ancêtre B. La comparaison des interactions possibles aux ancêtres A et B, permet alors d'identifier l'acquisition ou la perte d'interactions associées à chaque gène ancestral. Dans la figure, il y a acquisition de l'interaction bleue à l'ancêtre A.

7.1 Construction d'un jeu d'Interactions Protéine-Protéine

À la différence de l'identification des modules, l'étude des phénotypes associés à un ensemble de séquences nécessite un *a priori*. En effet, pour associer des phénotypes à des séquences, ici des interactions Protéine-Protéine, il est nécessaire d'utiliser les connaissances préalables. Ces connaissances préalables sont issues soit de manipulations expérimentales, soit d'outils de prédiction. Elles sont généralement stockées dans différentes bases de données. Notre but est ici d'extraire des bases de données les connaissances sur les interactions physiques entre deux protéines, impliquant une (ou plusieurs) ADAMTS-TSL humaine(s) pour lesquelles il existe une preuve expérimentale. Afin de récupérer des données d'interaction Protéine-Protéine, nous avons exploré deux stratégies : 1) les inter-

actions Protéine-Protéine présentes dans différentes bases de données et récupérables via l'interface `PSICQUIC` [Ara+11] (voir Section 1.2.3.3), et 2) les diverses interactions identifiées dans la littérature et non présentes dans PSICQUIC (e.g., interaction impliquant une enzyme et son substrat). Nous avons ensuite constitué un jeu de données *Synthèse-PPI* regroupant les données issues de ces deux sources distinctes d'information (Figure 7.2).

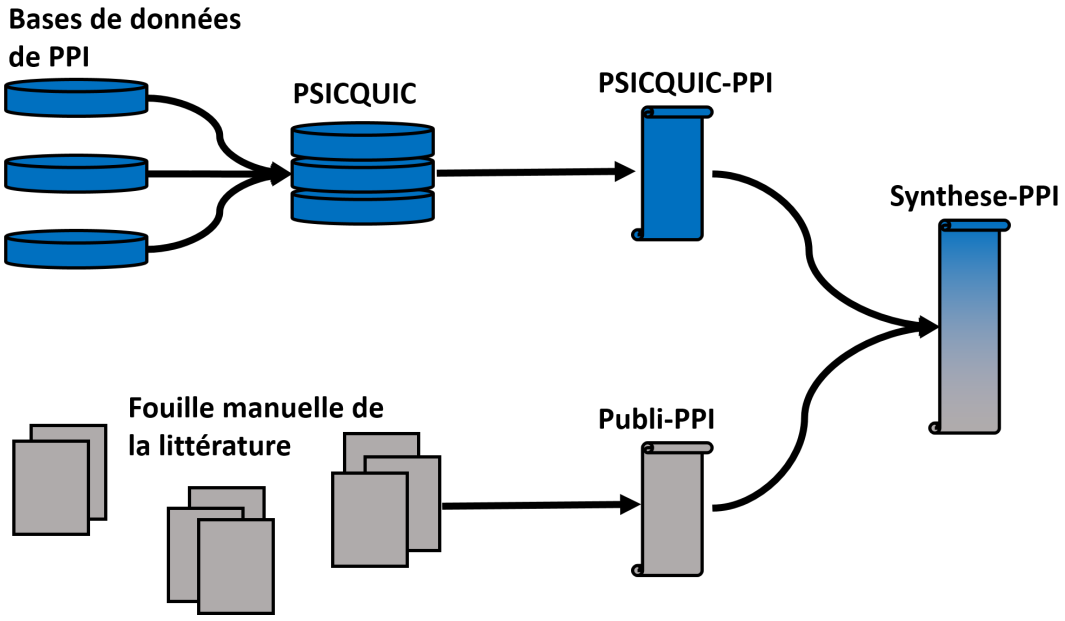

FIGURE 7.2 – **Synthèse des Interactions Protéine-Protéine issues de PSICQUIC et de la bibliographie des ADAMTS-TSL.**
Les Interactions Protéine-Protéine impliquant les ADAMTS-TSL humaines récupérées dans différentes bases de données de PPI, via l'interface `PSICQUIC` [Ara+11], composent le jeu de PPI *PSICQUIC-PPI*. Les Interactions Protéine-Protéine issues d'une fouille manuelle de la littérature composent le jeu de PPI *Publi-PPI*. Le fichier *Synthese-PPI* représente l'union des deux jeux.

Notre étude des PPI porte spécifiquement sur les 26 ADAMTS-TSL humaines. Ce que nous voulons étudier, c'est la capacité d'une ADAMTS-TSL à interagir avec un partenaire, par un mécanisme d'interaction partagé par une protéine ADAMTS-TSL humaine. Le mécanisme sera supposé exister depuis l'ancêtre commun des deux ADAMTS-TSL.

7.1.1 Extraction de PSICQUIC : 447 PPI impliquant les 26 ADAMTS-TSL humaines

Afin de récupérer un jeu de données d'Interactions Protéine-Protéine (PPI) impliquant les ADAMTS-TSL humaines, notre première approche a consisté à interroger les différentes bases de données dédiées (Figure 7.3) via l'interface `PSICQUIC` [Ara+11] (*Proteomics Standards Initiative common query interface*, voir Section 1.2.3.3). Nous avons pour cela convertit les identifiants RefSeq des 386 ADAMTS et 226 ADAMTSL (issus de dataset-708) en identifiants UniprotKB. Nous avons ainsi obtenu 494 identifiants UniprotKB, représentant différentes isoformes des ADAMTS-TSL des 9 espèces. Ce sont ces 494 identifiants UniprotKB qui ont ensuite été utilisés pour interroger `PSICQUIC` (le 01/08/21), nous permettant d'obtenir 2766 PPI (impliquant 577 partenaires d'interactions distincts). Nous avons ensuite retenu les 447 PPI soutenues non redondantes et impliquant un des 26 gènes ADAMTS-TSL humains et une autre protéine humaine, de manière à construire le jeu de PPI *PSICQUIC-PPI*.

Isoformes et PPI : Dans les bases de données, les PPI sont associées à un identifiant UniprotKB spécifique qui représente une isoforme spécifique. Les différentes isoformes d'un gène peuvent être capables d'effectuer des PPI différentes. Les annotations de PPI ne sont cependant pas assez précises pour distinguer quelle isoforme interagit ou non, les PPI étant généralement associées par défaut à l'isoforme de « référence ». Nous avons choisi ici d'interroger les bases de données pour toutes les isoformes que nous possédons (jeu dataset-708), avant d'associer les PPI obtenues à leur gène.

7.1.2 Substrats et recherche manuelle : 296 PPI impliquant les 26 ADAMTS-TSL humaines

Nous avons observé qu'un certain nombre de substrats et d'interactants des ADAMTS-TSL récemment identifiés n'étaient pas encore recensés dans les bases de données interrogées. Nous avons alors choisi de compléter manuellement notre jeu de données initial à partir de données issues de recherches bibliographiques. Pour ceci, nous avons utilisé une quinzaine d'articles présentant des PPI et des substrats des ADAMTS-TSL : [Kel+15 ; SH21b ; Col+19 ; Pi+15 ; Bek+16 ; Led+21 ; Sch+18 ; Wan+19a ; Som+03 ; Wan+19b ; Moh+21b ; Fon+21 ; San20 ; Yam+14 ; Nan+19]. Nous avons ainsi recensé 296 PPI impliquant les 26 protéines ADAMTS-TSL humaines, de manière à construire le jeu de PPI *Publi-PPI*.

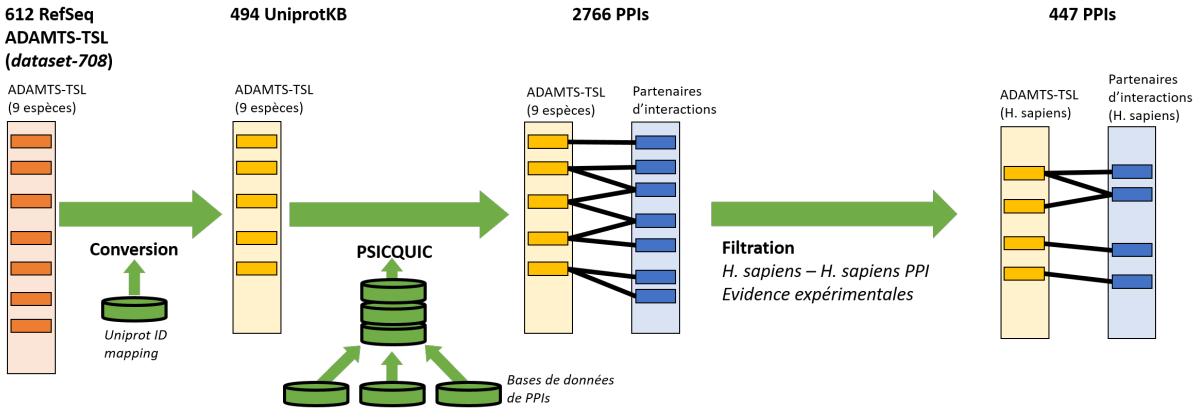

FIGURE 7.3 – **Récupération des Interactions Protéine-Protéine humaines grâce à l'interface PSICQUIC.**
Les identifiants RefSeq des 612 ADAMTS-TSL du jeu de séquences dataset-708 sont convertis en 494 identifiants UniprotKB. Interroger PSICQUIC à partir de ces identifiants UniprotKB nous permet de récupérer 2766 interactions Protéine-Protéine impliquant des ADAMTS-TSL. Filtrer les interactions impliquant uniquement deux partenaires humains, et basées sur des évidences expérimentales, permet de construire un jeu de 447 interactions impliquant les 26 ADAMTS-TSL humaines, dénommé *PSICQUIC-PPI*.

7.1.3 Synthèse des PPI des ADAMTS-TSL humaines

Le regroupement des 743 PPI (447 de *PSICQUIC-PPI* et 296 de *Publi-PPI*) issues de ces deux sources distinctes nous a permis de constituer le jeu de données de PPI *Synthèse-PPI*, qui regroupe de manière non redondante 720 PPI humaines documentées. Ces 720 PPI impliquent toutes une des 26 ADAMTS-TSL humaine et un des 471 partenaires humains identifiés. Parmi ces 471 partenaires, 119 interagissent avec au moins deux ADAMTS-TSL différentes (Figure 7.4), et peuvent alors être partagés depuis un ancêtre commun correspondant à une duplication paralogue ou être issus d'une convergence évolutive.

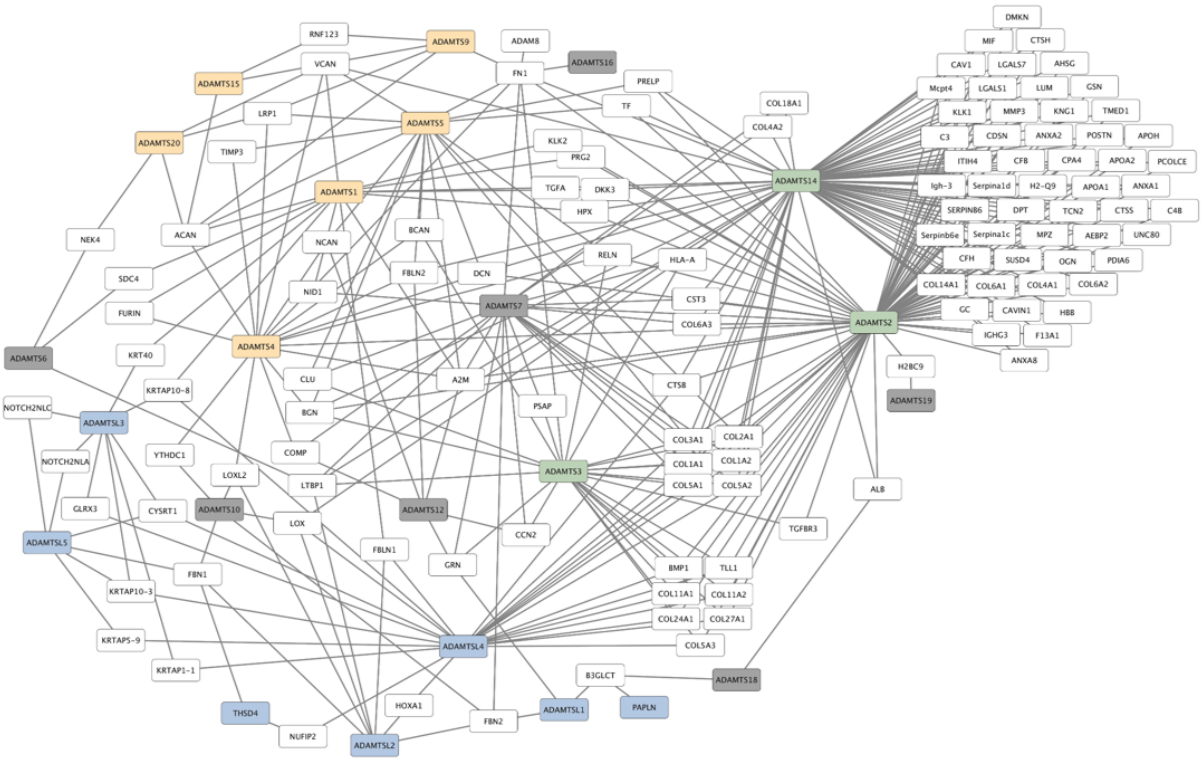

FIGURE 7.4 – **Graphe des Interactions Protéine-Protéine partagées par les 26 ADAMTS-TSL humaines.**
Graphe des 119 interactions partagées par au moins deux des 26 ADAMTS-TSL humaines, du jeu de données de PPI *Synthèse-PPI*. La visualisation a été réalisée avec le logiciel `Cytoscape` [Sha+03].

7.2 Annoter les séquences ADAMTS-TSL avec leurs interactions

Le jeu de données *Synthèse-PPI* constitue une liste d'observations expérimentales, où chacune des 720 PPI décrit une interaction entre deux partenaires : le premier partenaire est une des 26 protéines ADAMTS-TSL humaines, et le deuxième partenaire est l'une des 471 protéines identifiées comme interagissant avec elle. Notre but est maintenant d'utiliser les observations que sont ces PPI pour les interpréter comme traits phénotypiques des ADAMTS-TSL humaines (Figure 7.5). Nous considérons donc qu'une ADAMTS-TSL peut interagir ou ne pas interagir avec chacune des protéines partenaires identifiées. La capacité d'une ADAMTS-TSL A à interagir avec un partenaire B est nommée *interaction avec B* et peut avoir été observée par une PPI impliquant les partenaires A et B. Nous décrivons 471 interactions, chaque interaction correspond à la capacité à interagir avec un des 471 partenaires.

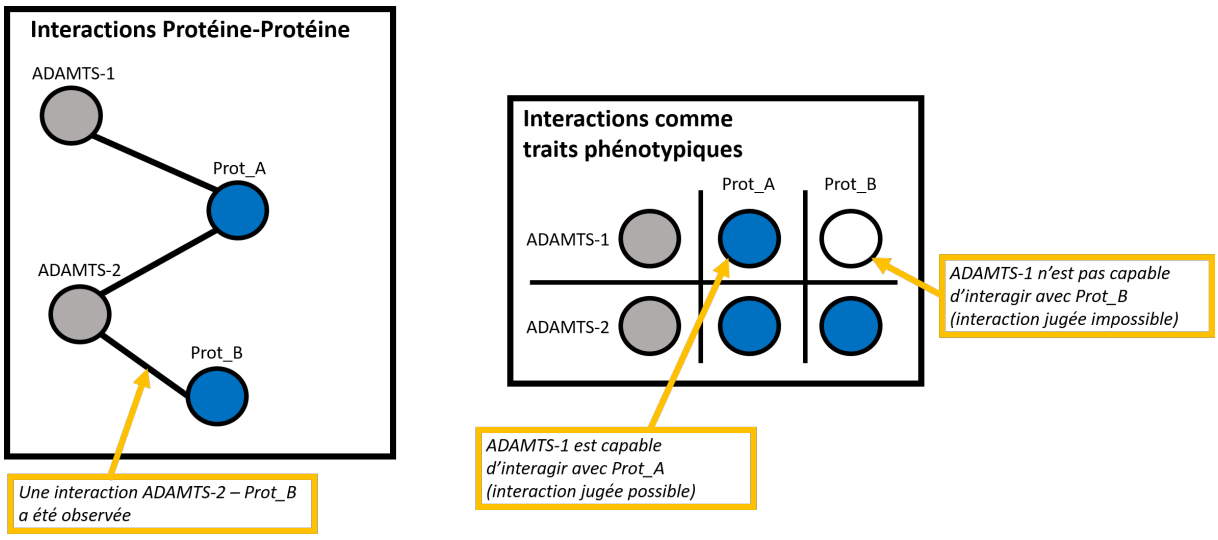

FIGURE 7.5 – **Utilisation des PPI pour associer les possibilités d'interactions comme phénotypes.**
Une Interaction Protéine-Protéine (PPI) décrit une interaction observée entre deux partenaires (e.g., ADAMTS-1 interagit avec Prot_A). Les PPI sont utilisées comme observations pour associer à chaque protéine d'intérêt (en gris) sa capacité à interagir ou non avec les différents partenaires identifiés (en bleu). Un trait phénotypique d'une protéine d'intérêt est alors décrit comme sa capacité à interagir (en bleu) ou non (en blanc) avec les partenaires identifiés chez elle. Une interaction est alors jugée comme possible ou impossible.

Une difficulté ici est que nous possédons uniquement des observations sur la capacité d'une protéine ADAMTS-TSL à interagir avec un partenaire, mais pas (ou très peu) d'observations sur l'incapacité de celle-ci à interagir avec un autre partenaire. En effet, l'absence d'une PPI dans une base de données ou une publication ne signifie pas qu'elle n'est pas possible, elle n'a juste pas été observée. À noter que de nouvelles PPI qui seront découvertes ultérieurement pourront être implémentées dans notre modèle et enrichir nos prédictions.

Nous allons inférer l'évolution de ces 471 interactions au sein de l'arbre des gènes (arbre de référence, voir Chapitre 5). Pour ceci, nous utilisons l'outil `PastML` [Ish+19] afin d'inférer le scénario ancestral de chacune de ces 471 interactions. Un scénario ancestral est l'estimation de l'évolution d'un caractère le long d'une phylogénie, où pour chaque ancêtre le caractère est estimé présent ou non (voir Section 7.3). Afin d'utiliser `PastML`, nous avons besoin d'associer toutes les feuilles de l'arbre des gènes avec un *état* pour chaque caractère que nous allons étudier. Nous considérons comme caractère pour l'inférence avec `PastML` la capacité d'interaction d'une ADAMTS-TSL (i.e., une feuille de l'arbre des gènes) avec une autre protéine (i.e., une des 471 partenaires). Nous décrivons le caractère par un couple de protéines : ADAMTS-TSL - partenaire d'interaction. Ce couple est alors associé à l'état de l'interaction. Cet état peut-être : *l'interaction est possible* (1), ou *l'interaction est impossible* (0). `PastML` permet également de considérer un caractère comme *non déterminé* (?), `PastML` se chargera alors de déterminer l'état (Figure 7.6). Dans notre contexte, une interaction considérée comme *non déterminée* sera alors inférée comme *possible* ou *impossible*.

Une approche logique serait d'estimer toutes les interactions présentes dans *Synthèse-PPI* comme *possibles*, et toutes les autres comme *non déterminées* (Figure 7.7). De ce fait, une PPI non observée est alors inférée par `PastML`. Cependant, en l'absence d'interactions annotées comme *impossible* (0), tous les couples seront alors associés avec un (1) ou (?). Les (?) ne pourront alors qu'être inféré en (1). En effet, si on ne possède que des interactions *possibles* et des interactions dont l'état est inconnu, sans aucune interaction *impossible*, le scénario le plus probable est que les toutes les interactions soient *possibles*. Il est donc nécessaire d'estimer certaines absences de données d'interaction comme des interactions *impossibles*.

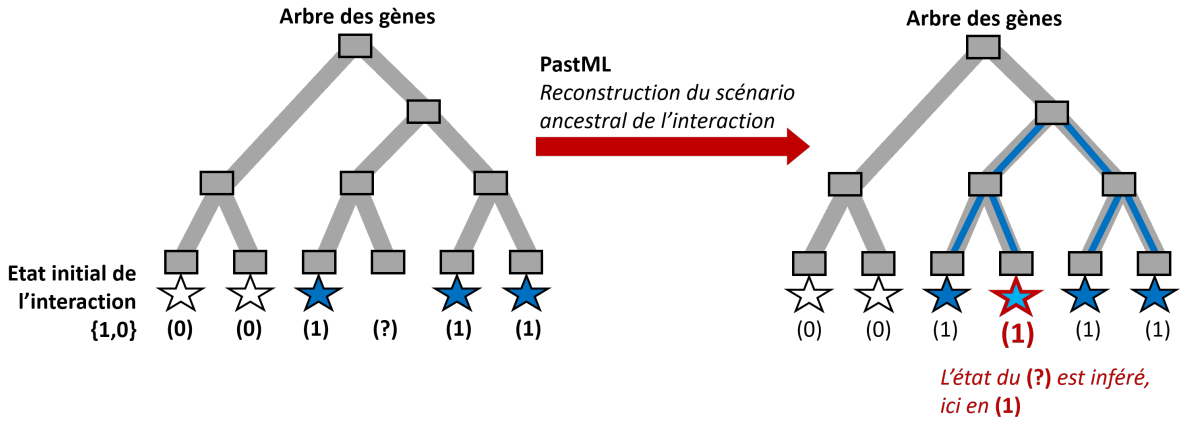

FIGURE 7.6 – **Une interaction comme caractère phénotypique discret analysé par PastML : (1) (0) (?).**
Pour une interaction donnée (ici l'étoile bleue), toutes les feuilles de l'arbre des gènes sont associées à un *état* parmi *possible* (1), *impossible* (0), et *non déterminée* (?), selon que le gène (la feuille) est jugé capable ou non de réaliser l'interaction. PastML inférera le scénario ancestral de cette interaction, inférant également l'état des interactions *non déterminées* (en rouge sur la figure).

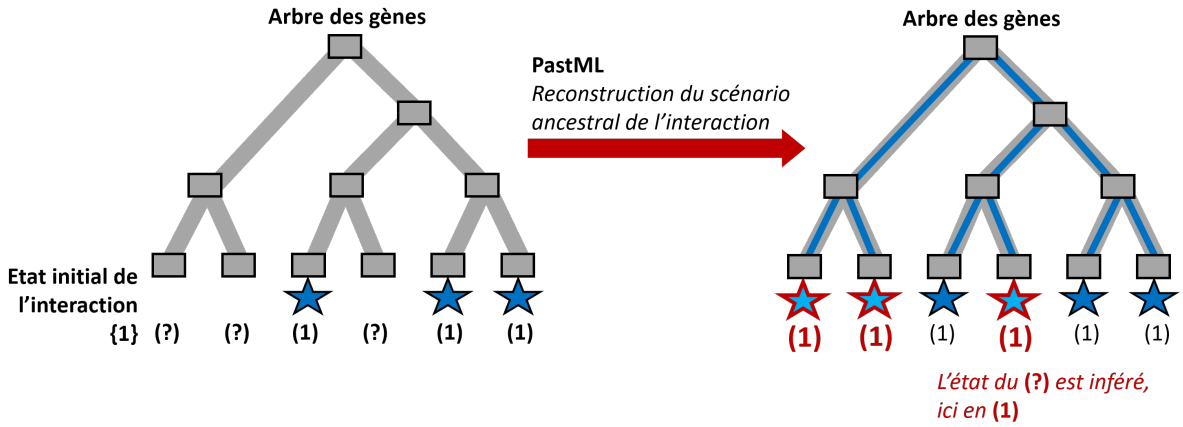

FIGURE 7.7 – **L'importance des interactions impossibles.**
Pour une interaction donnée (ici l'étoile bleue), toutes les feuilles de l'arbre des gènes sont associées à un état parmi *possible* (1), et *non déterminée* (?), en fonction de si l'interaction a été observée ou non. PastML inférera le scénario ancestral de cette interaction, inférant également l'état des possibles interactions *non déterminées* (en rouge), qui ne pourra être que *possible*.

Notre objectif est d'associer à chaque couple d'interaction : ADAMTS-TSL - partenaire d'interaction, un état, parmi : *possible* (1), *impossible* (0) ou *non déterminé* (?). Les 214 feuilles ADAMTS-TSL seront alors associées à un état pour chacune des 471 interactions possibles, ce qui représente une matrice de 100 794 couples d'interactions (214×471). Pour annoter ces 100 794 couples, tout en estimant l'absence de données, nous avons exploré deux approches concurrentes : les interactions sont partagées par les protéines orthologues (Section 7.2.1), et seules les interactions humaines sont considérées comme observées (Section 7.2.2). En dernier lieu, c'est cette seconde approche (Section 7.2.2) qui sera retenue.

Chacune des 26 ADAMTS-TSL humaines est associée à un état parmi : *possible* (1) et *impossible* (0) pour les 471 interactions recensées. Si une PPI a été observée, l'interaction est considérée *possible*, mais si la PPI n'a pas été observée, l'interaction est alors considérée comme *impossible*. L'hypothèse est que si une interaction n'a pas été observée chez une ADAMTS-TSL, alors que pour au moins une autre ADAMTS-TSL elle a été observée (et donc étudiée), alors il y a plus de chance qu'elle soit impossible. Il est possible que cette interaction soit possible, cependant tant qu'elle n'a pas été observée, elle sera considérée comme impossible jusqu'à qu'elle soit observée. Notre modèle est ainsi amené à évoluer en fonction des découvertes de nouvelles interactions.

7.2.1 Première approche : propager systématiquement les interactions aux orthologues

Des protéines orthologues sont généralement considérées plus similaires fonctionnellement que des protéines paralogues [TKL97 ; DB07 ; GK13]. Notre première hypothèse se base sur cette hypothèse et consiste à propager les interactions des séquences humaines à leurs séquences orthologues (Figure 7.8). Nous utilisons alors l'arbre de gènes et les événements Gène-Espèce issus de la réconciliation Module-Gène-Espèce. Cette réconciliation associe à chaque nœud ancestral de l'arbre des gènes un événement parmi : *spéciation*, *duplication de gène*. Nous décrivons ici les *protéines orthologues* à une protéine humaine comme des protéines séparées de la protéine humaine uniquement par des spéciations. Ces protéines orthologues sont considérées comme similaires fonctionnellement à la protéine humaine et sont associées aux mêmes PPI que la protéine humaine (en amont de l'inférence `PastML`).

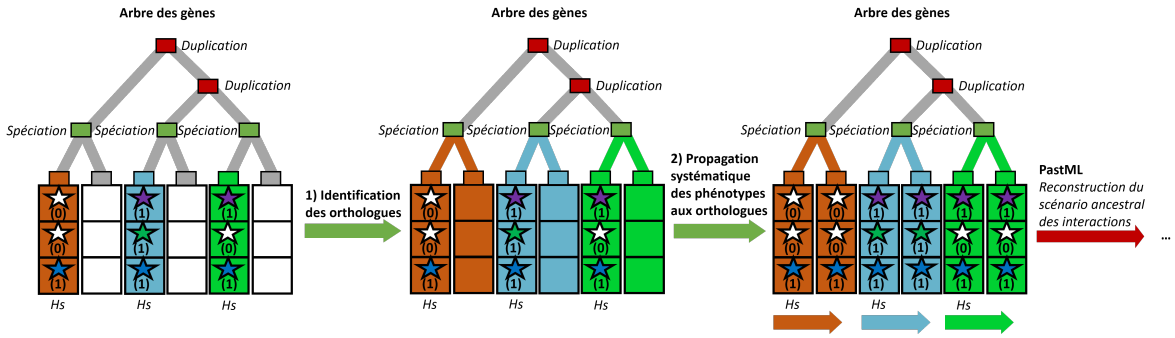

FIGURE 7.8 – **Propagation systématique des phénotypes aux orthologues.**
Initialement, seules les séquences des protéines humaines (*Hs*) sont associées avec des interactions. Les trois interactions considérées dans cet exemple, sont représentées par des étoiles : remplie si l'interaction est possible (1), vide si elle est considérée comme impossible (0). Chaque nœud ancestral de l'arbre des gènes est associé à un événement parmi : *spéciation* (rectangle vert) et *duplication de gène* (rectangle rouge, à gauche), les orthologues des protéines humaines sont identifiées sur la base de ces évènements (au milieu), puis les interactions des protéines humaines sont propagées à leurs orthologues (à droite).

7.2.2 Seconde approche : laisser PastML inférer les interactions chez les orthologues

L'hypothèse d'orthologie (Section 7.2.1) se base sur un *a priori* important : les orthologues partagent leurs fonctions. Cependant, une étude de Stamboulian et al. [Sta+20] a montré que la similarité de fonctions n'est pas propre aux orthologues ou aux paralogues, mais dépend principalement de la divergence entre ces séquences (i.e., longueur de branches dans une phylogénie), la méthode `PastML` considère cette divergence des séquences. C'est pourquoi notre seconde hypothèse consiste à laisser un plus grand degré de liberté à `PastML`, en utilisant en entrée uniquement les interactions des protéines humaines. Ceci a pour but de laisser `PastML` inférer les interactions des protéines non-humaines, en s'appuyant sur la répartition des interactions humaines, et surtout sur la longueur de branches de l'arbre des gènes (Figure 7.9).

Comme ce sont les interactions humaines qui nous intéressent, nous préférons ne considérer aucun *a priori* pour inférer leur présence ou non chez les protéines des autres espèces. En laissant `PastML` les inférer, nous considérons les longueurs de branches et donc l'hypothèse suivante : plus une séquence évolue rapidement et diverge, moins l'interaction a de chance d'être conservée. À l'inverse, moins elle diverge, plus l'interaction a de chance d'être conservée. Pour plus de détail sur l'inférence des scénarios ancestraux, et donc de ces interactions *non déterminées* voir la Section 7.3. En appliquant cette approche à notre arbre

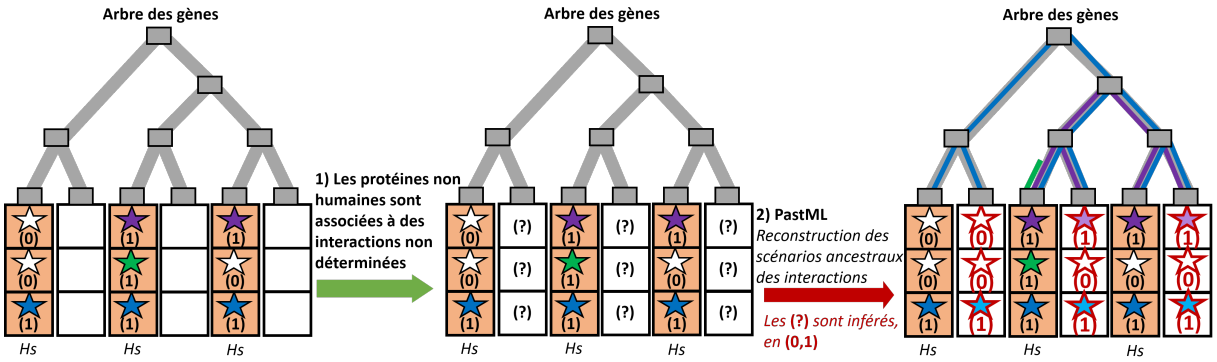

FIGURE 7.9 – **Inférence des phénotypes par PastML.**
Initialement, seules les séquences des protéines humaines (*Hs*) sont associées avec des interactions. Les trois interactions considérées dans cet exemple, sont représentées par des étoiles : violette/verte/bleue si l'interaction est possible (1), blanche si elle est considérée comme impossible (0). **1)** Les interactions des protéines non humaines sont considérées comme non déterminées (?) .**2)** Les interactions non déterminées sont inférées par `PastML`, sur la base de leur répartition dans l'arbre, et de la longueur des branches de la phylogénie.

de 214 séquences, `PastML` infére à partir des 26 ADAMTS-TSL humaines, pour chacune des 188 protéines homologues ADAMTS-TSL non-humaines, la possibilité/impossibilité d'interagir avec chacune des 471 protéines connues pour être partenaire d'au moins une ADAMTS-TSL humaine (Figure 7.10).

En conclusion, le choix a été fait de nous focaliser sur les interactions telles qu'elles existent chez l'humain (seconde approche, Section 7). Pour ceci, nous avons uniquement annoté les 26 ADAMTS-TSL humaines avec les interactions recensées dans notre jeu de PPI, en considérant une PPI non observée comme impossible. Sans *a priori*, c'est ensuite `PastML` qui infère l'état des interactions chez les protéines non-humaines. Ces inférences chez les homologues reposent sur un modèle évolutif qui considère la répartition d'une interaction, et la longueur des branches (degrés de divergence des séquences).

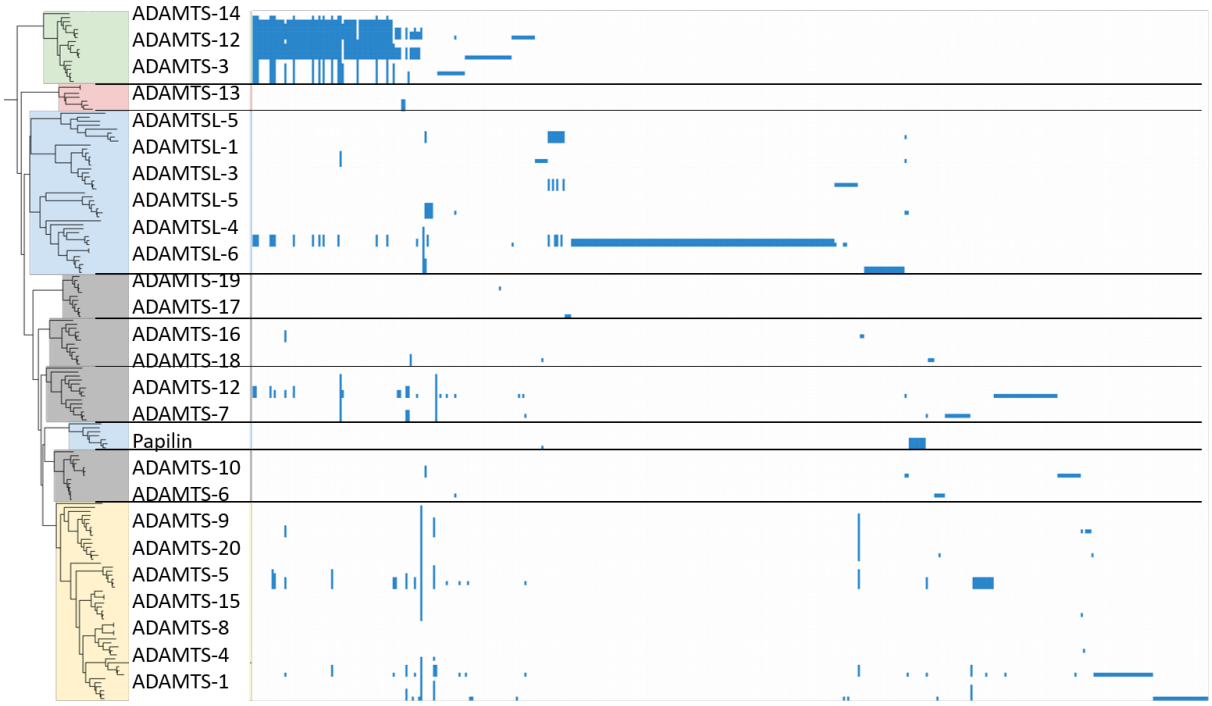

FIGURE 7.10 – **Présence des 471 interactions pour les 214 ADAMTS-TSL d'après PastML.** Les 720 PPI du jeu de données *Synthèse-PPI* nous ont permis d'annoter les 26 ADAMTS-TSL humaines de l'arbre des 214 gènes. Pour une ADAMTS-TSL humaine, chacun des 471 interactants est alors associé à la valeur (1) quand l'interaction est jugée possible ou (0) quand l'interaction est jugée impossible. Les 188 autres protéines ont toutes leurs interactions associées avec un état inconnu (?), de manière à être inférées par `PastML`. Les états inconnus (?) sont alors estimés comme possible (1) ou impossible (0) en fonction des scénarios ancestraux. Chaque feuille de l'arbre correspond à une protéine, et les colonnes correspondent aux 471 interctants. Un carré bleu représente alors une interaction possible, et l'absence d'un carré bleu, une interaction impossible.

7.3 Évolution des Interactions Protéine-Protéine au sein de l'évolution des gènes ADAMTS-TSL

Le partage de phénotypes par des protéines homologues peut être dû à une origine évolutive commune, ou à une acquisition indépendante de phénotypes similaires (i.e., convergence évolutive/homoplasie). Notre hypothèse est qu'un phénotype partagé par des protéines homologues, mais résultant d'origines évolutives différentes et convergentes, est associé à des mécanismes et des résidus fonctionnels différents. Étudier l'évolution des phénotypes permet de distinguer un phénotype partagé, issu d'une origine évolutive commune, d'un phénotype convergent qui aurait été acquis de manière indépendante. Notre objectif est d'inférer l'évolution des interactions au cours de l'évolution des gènes ADAMTS-TSL (Figure 7.11), c'est-à-dire inférer pour chacune des 471 interactions (capacité d'interagir avec l'un des 471 partenaires identifiés des ADAMTS-TSL) sa présence/absence à chaque gène ancestral (Section 7.3.1) de l'arbre de référence (inféré préalablement en Section 5.3.1). Nous estimons ainsi les gains/pertes d'interactions au cours de l'évolution des gènes ADAMTS-TSL (Section 7.3.2). Dans une dernière étape, nous croiserons ces phénotypes ancestraux avec les compositions ancestrales en modules conservés.

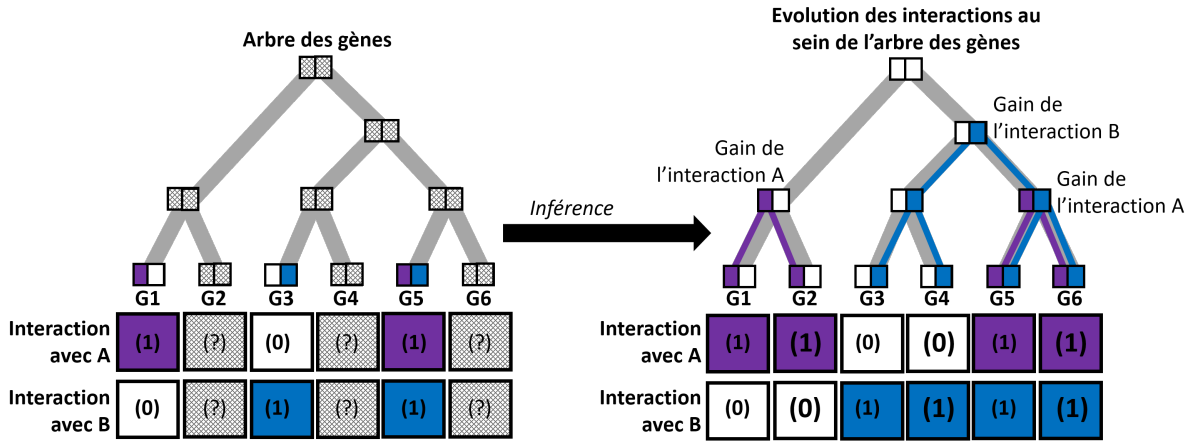

FIGURE 7.11 – **Inférer l'évolution des interactions au sein de l'arbre des gènes et identifier les nœuds d'acquisition des interactions.**
Sur l'arbre des gènes p1, p2, p3, p4, p5, p6, seules les feuilles p1, p3, p5 sont annotées avec la possibilité (1) et impossibilité (0) de réaliser les interactions avec les partenaires A et B (violet et bleu, respectivement). La possibilité est représentée en couleur, l'impossibilité en blanc. Les feuilles (p2, p4, p6) et les nœuds ancestraux dont les interactions ne sont pas déterminées (?) sont représentés par des carrés gris. La reconstruction des deux scénarios ancestraux par `PastML` (un scénario par interaction) permet de résoudre les interactions qui n'étaient pas déterminées, qui sont alors associées à des (1) ou des (0). Étudiant la présence des interactions des différents gènes et de leurs ancêtres, il est alors possible d'inférer l'acquisition (gain) des interactions. L'inférence phylogénétique effectuée par `PastML` fait apparaître que l'interaction A est partagée par les protéines p1, p2 et p5, p6, mais est issue de deux acquisitions indépendantes. De plus, l'inférence phylogénétique effectuée par `PastML` fait apparaître que l'interaction B est partagée par les protéines p3, p4, p5 et p6, et est issue d'une unique acquisition.

7.3.1 Interactions ancestrales des ADAMTS-TSL

Afin d'inférer l'évolution des phénotypes, nous avons utilisé le programme `PastML` [Ish+19] qui permet de reconstruire les scénarios ancestraux de caractères sur un arbre phylogénétique dont les feuilles sont annotées. Pour un caractère donné, `PastML` utilise l'annotation des différentes feuilles (décrivant l'état du caractère à chaque feuille) et reconstruit un scénario ancestral sur l'arbre des gènes par maximum de vraisemblance et parcimonie, tout en considérant la topologie et les longueurs des branches de l'arbre. Dans le cadre de notre étude des interactions des ADAMTS-TSL, un caractère est une interaction (la capacité d'interagir avec l'un des 471 partenaires d'interactions identifiés), ce qui signifie que pour les 471 interactions identifiées, nous avons inféré un scénario ancestral. Une étape préalable consiste à annoter les différentes feuilles de l'arbre avec l'état des différentes interactions. Nous avons abordé cette étape dans un des paragraphes précédents (voir Section 7.2.2) : seules les interactions des 26 ADAMTS-TSL humaines sont annotées, les interactions des 188 autres homologues sont laissées *non déterminées*, de manière à laisser `PastML` les inférer. Inférer le scénario ancestral d'une interaction consiste à inférer la présence ou l'absence de l'interaction aux différents gènes ancestraux de l'arbre des gènes (Figure 7.12).

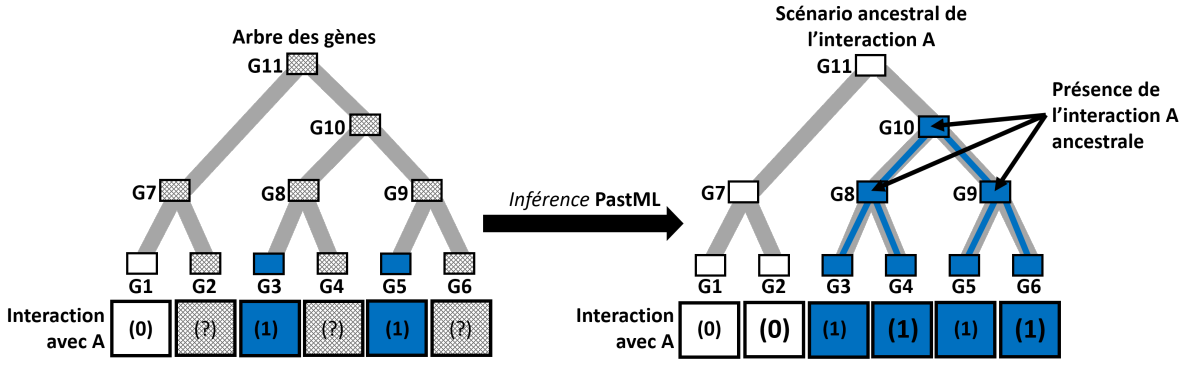

FIGURE 7.12 – **Inférer le scénario ancestral d'une interaction au sein de l'arbre des gènes.** Inférer avec `PastML` le scénario ancestral de l'interaction A, sur la base de l'annotation des feuilles p1, p3 et p5, permet d'associer la présence de l'interaction A (en bleu) aux gènes ancestraux a3, a4, a5 et aux feuilles p4, p6. Cela permet également d'associer l'absence de l'interaction A (en blanc) aux gènes ancestraux a1, a2 et à la feuille p2.

La Figure 7.13 montre les scénarios ancestraux inférés par `PastML` pour les 471 interactions du jeu de données. Enfin, il faut noter que `PastML` propose plusieurs scénarios. Nous utilisons uniquement le plus probable, sur la base des probabilités marginales qui

sont indiquées. Nous pouvons ainsi décrire un gène de l'arbre des gènes par la liste des interactions qui y sont présentes.

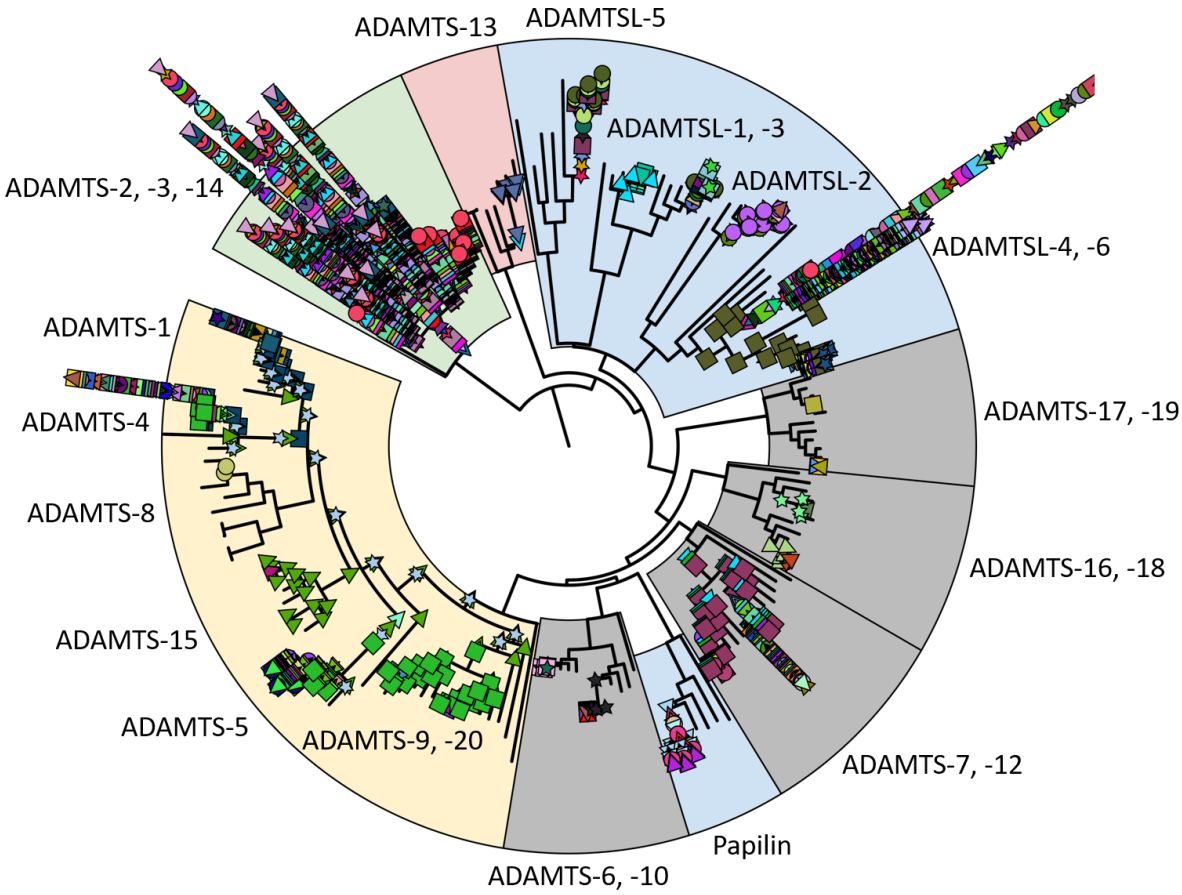

FIGURE 7.13 – **Évolution des 471 interactions des protéines ADAMTS-TSL.**
La présence de chacune des 471 interactions est indiquée pour chaque nœud de l'arbre des gènes ADAMTS-TSL (le groupe externe n'est pas représenté). Chacune des interactions est représentée par une combinaison forme/couleur spécifique. On notera par exemple la disparité de répartition des interactions, le clade ADAMTS-2, -3, -14 a un très grand nombre d'interactions qui remontent à l'ancêtre du clade. À l'inverse, la protéine ADAMTS-8 est annotée avec une unique interaction (représentée par un cercle jaune) qui ne remonte pas dans l'arbre. Les interactions remontent au maximum aux ancêtres des sous-groupes fonctionnels (préalablement identifiés comme robustes, Chapitre 5), c'est par exemple le cas de l'interaction avec ACAN (étoile grise) qui remontent à l'ancêtre des hyalectanases (sous-groupe jaune).

7.3.2 Gains et pertes d'interactions au cours de l'évolution des gènes ADAMTS-TSL

Chacun des 427 gènes (213 ancestraux et 214 actuels) de l'arbre de référence ADAMTS-TSL est décrit par une liste des interactions qui y sont possibles. D'une manière similaire à notre traitement de l'évolution des modules (Section 6.4), nous pouvons alors utiliser la topologie de l'arbre de référence afin d'observer les changements en terme d'interactions possibles, d'un gène ancestral à son descendant. Pour ceci, nous comparons les interactions d'un gène et de son ancêtre direct. Nous identifions trois situations possibles : 1) une interaction avec un partenaire P est possible chez un gène (actuel ou ancestral) et chez son ancêtre, 2) une interaction est possible chez un gène, mais impossible chez son ancêtre (*interaction gagnée*), et 3) une interaction est impossible chez un gène, mais possible chez son ancêtre (*interaction perdue*). Un gène (nœud de l'arbre des gènes) est alors décrit par un ensemble d'états pour les interactions : les interactions présentes (possibles), les interactions gagnées et les interactions perdues (Figure 7.14).

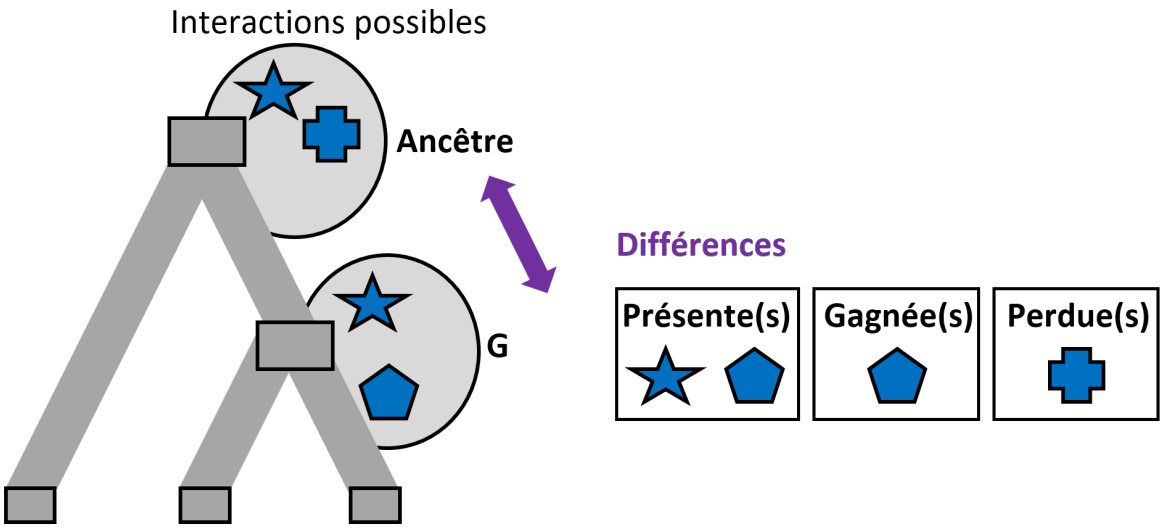

FIGURE 7.14 – **Évolution des interactions possibles.**
Chaque gène (nœud de l'arbre des gènes) est associé à la liste de ses interactions possibles. L'évolution des interactions possibles est modélisée par les gains et les pertes d'interactions possibles entre un gène et son ancêtre.

Dans nos travaux, nous allons nous focaliser uniquement sur les interactions gagnées, dans le but d'identifier et d'étudier les gènes ancestraux où un phénotype, ici une interaction, est gagnée au cours de l'évolution des gènes ADAMTS-TSL. Nous avons identifié

48 gènes ancestraux (parmi 213) où au moins une interaction est gagnée. Plusieurs interactions peuvent être gagnées par un même gène ancestral, empreinte d'une importante divergence fonctionnelle de ce gène. Nous allons par la suite (Chapitre 8) étudier les modules qui sont gagnés en même temps que les interactions (i.e., *co-apparition module(s)-interaction(s)*) afin de proposer ces modules comme des régions associées à ces interactions.

Conclusion

Les ADAMTS-TSL se distinguent fonctionnellement par leur capacité à interagir avec différents partenaires moléculaires. Nous nous sommes focalisés sur leurs interactions physiques avec d'autres protéines. Nous avons identifié un jeu d'Interactions Protéine-Protéine (PPI) appelé *Synthèse-PPI*, qui regroupe 720 interactions Protéine-Protéine impliquant les 26 ADAMTS-TSL humaines et 471 partenaires d'interactions. Ce jeu de PPI que nous avons voulu le plus exhaustif possible, est issu de différentes bases de données et d'une analyse bibliographique. Nous avons alors défini 471 traits phénotypiques (les capacités d'interactions), qui correspondent à la capacité d'une des 26 ADAMTS-TSL humaine à interagir ou non avec l'un des 471 partenaires. Chacun de ces 471 traits phénotypiques (capacité d'interagir avec l'un des partenaires) est tout d'abord associé aux 26 protéines ADAMTS-TSL humaines, à partir des connaissances du jeu *Synthèse-PPI*. Pour chacune de ces 471 interactions, nous avons inféré son scénario ancestral au sein de l'arbre de référence des 214 séquences ADAMTS-TSL, de manière à associer à chacun des 213 gènes ancestraux et à chacune des 188 feuilles non annotées (homologues non humains) de l'arbre de référence, sa capacité à interagir ou non avec chacun des 471 partenaires. Ceci nous a permis d'identifier 48 gènes ancestraux où au moins une interaction y a été acquise. Cette acquisition représente l'empreinte d'une divergence fonctionnelle à ce moment de l'évolution des ADAMTS-TSL. Notre objectif final est d'étudier de manière conjointe ces gains d'interactions avec les gains de modules, de manière à caractériser des régions fonctionnelles apparues conjointement à ces interactions, et donc en mesures à être impliquées dans ces interactions. Ceci fait l'objet du chapitre suivant (Chapitre 8).

Chapitre 8

Joindre l'évolution des modules et des interactions protéine-protéine au sein de l'arbre des gènes ADAMTS-TSL

Afin d'identifier de nouveaux modules fonctionnels des protéines ADAMTS-TSL, nous proposons d'étendre la méthode phylogénomique aux modules de séquences conservés. La phylogénomique étudie l'évolution de protéines et de leurs fonctions, dans le but de propager ces fonctions dans une phylogénie, et ainsi de prédire de manière efficace la fonction de protéines inconnues [Eis98 ; Eng+05 ; Yon+20]. Ces méthodes associent des fonctions à des séquences complètes, notre objectif au cours de cette thèse est d'associer des fonctions à des régions des séquences, sur la base de leur évolution, permettant ainsi l'extension des méthodes phylogénomiques à la prédiction de modules fonctionnels. L'hypothèse sous-jacente est la suivante : l'évolution des régions impliquées dans une fonction est corrélée à l'évolution de la fonction dans laquelle ces régions sont impliquées. Différentes études confirment cette hypothèse et associent des régions spécifiques de séquences à leurs fonctions [Gau+11 ; FG13 ; Lau+21] sur la base de leurs évolutions corrélées. Nous visons ici à étendre ce type de raisonnement à la totalité de la famille ADAMTS-TSL en étudiant de manière conjointe l'origine des modules conservés et des interactions protéine-protéine. C'est pourquoi nous avons inféré : l'évolution des gènes (Chapitre 5), l'évolution des compositions en modules (Chapitre 6) et l'évolution des interactions protéine-protéine (Chapitre 7). Il s'agit maintenant de les interpréter conjointement afin de caractériser les événements de *coapparitions module(s) - interaction(s)*, afin d'associer modules et interactions.

8.1 L'arbre des gènes comme modèle

L'évolution des modules (Chapitre 6) ainsi que l'évolution des phénotypes (ici, les interactions, Chapitre 7) ont toutes les deux été inférées en considérant l'arbre de référence des gènes ADAMTS-TSL (Chapitre 5). Ceci nous permet d'utiliser l'arbre de référence comme modèle, où chaque gène (i.e., nœud de l'arbre, actuel ou ancestral) est associé à la liste de ses modules présents, gagnés et perdus, ainsi que la liste de ses interactions présentes, gagnées et perdues (Figure 8.1). Cette intégration au sein d'un même arbre des gènes nous permet d'étudier conjointement l'évolution des modules et des interactions au cours de l'évolution des gènes ADAMTS-TSL.

Nous avons formalisé l'arbre-modèle de la manière suivante : soit $\mathcal{M}$ et $\mathcal{I}$ les modules et les interactions associées au jeu de gènes homologues. Pour chaque nœud n de l'arbre des gènes (qu'il soit un nœud feuille ou un nœud interne) nous associons à une liste d'interactions $\mathcal{I}_n$ et une liste de modules $\mathcal{M}_n$. La comparaison d'un nœud n avec son ancêtre a permet également de calculer les listes $\mathcal{M}_n^+$ et $\mathcal{M}_n^-$ des modules gagnés et perdus. Un module est ajouté à $\mathcal{M}_n^+$ s'il est présent dans $\mathcal{M}_n$ mais absent de $\mathcal{M}_a$. Un module est ajouté à $\mathcal{M}_n^-$ s'il est dans $\mathcal{M}_a$ mais absent de $\mathcal{M}_n$. Les listes $\mathcal{I}_n^+$ et $\mathcal{I}_n^-$ des interactions gagnées et perdues sont calculées de la même manière. La liste des modules gagnés au nœud n, $\mathcal{M}_n^+$ définit alors sa signature de modules.

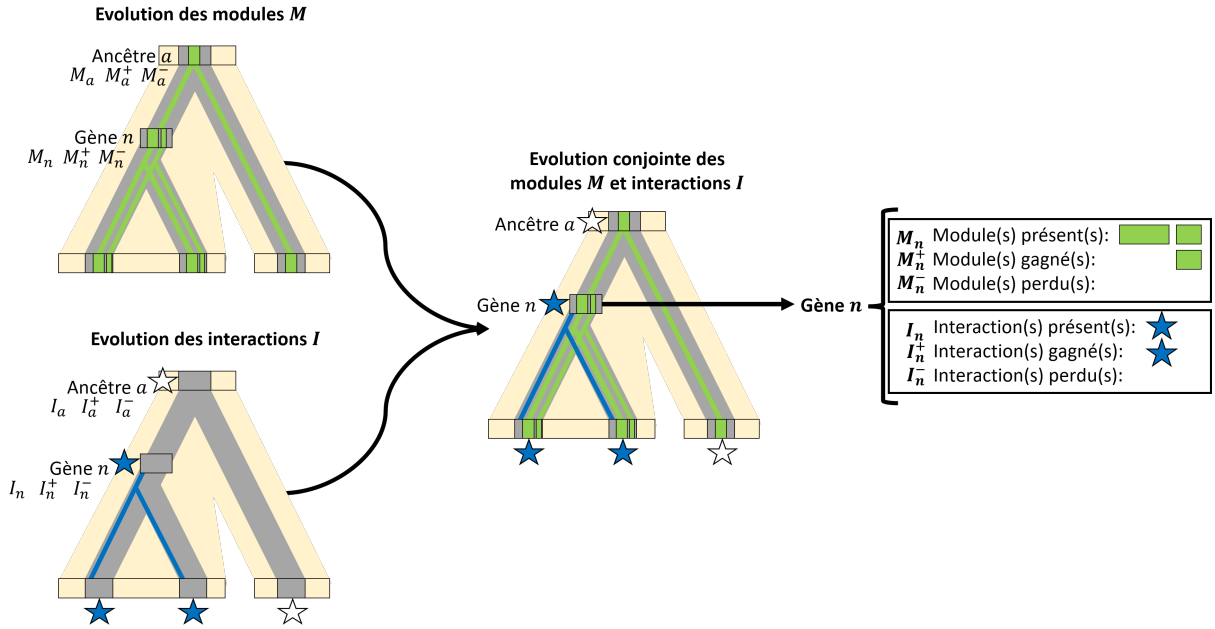

FIGURE 8.1 – **Joindre l'évolution des modules et l'évolution des interactions.**
Joindre l'évolution des compositions en modules et l'évolution des phénotypes permet d'obtenir pour chaque gène : ses modules présents, gagnés et perdus, ainsi que ses phénotypes présents, gagnés et perdus.

8.1.1 Une implémentation de la méthode

Nous proposons une implémentation de notre méthode (Figure 8.2). Le dépôt `https://github.com/OcMalde/PhyloCharMod_publ` regroupe les différents scripts `Python3` du *framework* `PhyloCharMod` (***Phylo****genetic* ***Char****acterization of* ***Mod****ules*) qui permet d'inférer les différentes histoires évolutives et de les regrouper conjointement au sein de l'arbre des gènes. Ce qui donne la possibilité d'appliquer ce *framework* `PhyloCharMod` à d'autres jeux de séquences homologues et d'autres types de phénotypes.

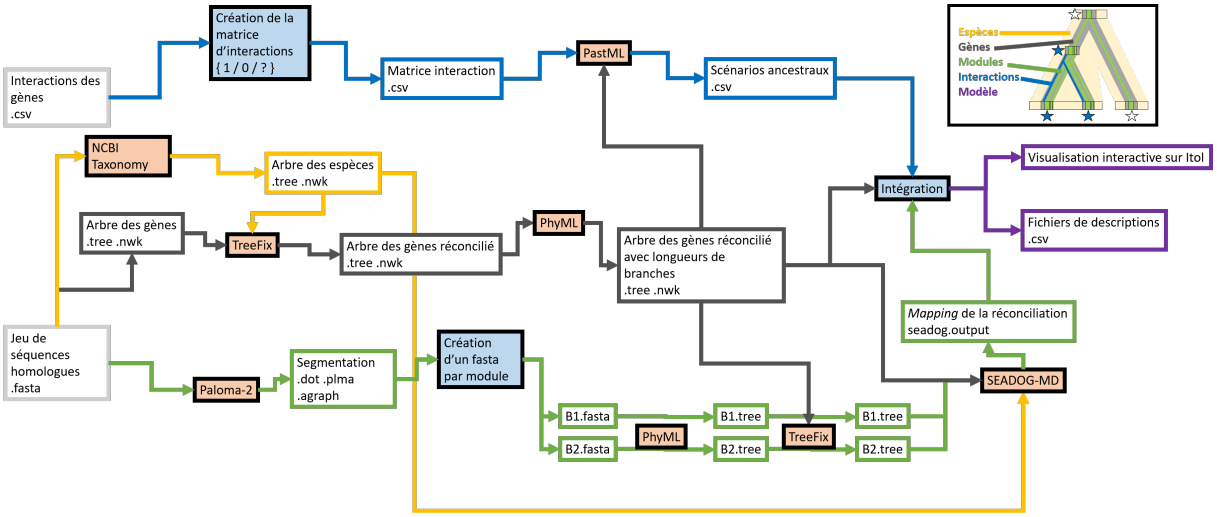

FIGURE 8.2 – **Schéma du Workflow PhyloCharMod.**
Représentation schématique des grandes étapes de notre implémentation. Celle-ci nécessite un jeu de séquences homologues (au format `fasta`), un arbre des gènes de ces séquences, ainsi qu'un fichier décrivant les interactions d'intérêts de ces gènes (au format `csv`), et construit l'évolution conjointe des modules et des interactions, décrite dans divers fichiers et dans une représentation interactive avec Itol (violet). Les outils existants utilisés sont représentés en orange, les étapes que nous avons implémentées en bleu, et les fichiers sont représentés par des boîtes sans couleur. Les flèches jaunes représentent les étapes impliquant les espèces, les flèches grises représentent les étapes impliquant les gènes, les flèches vertes celles impliquant les modules, et les bleues celles impliquant les interactions.

8.2 Représentation des évolutions jointes modules-interactions

Le regroupement des différents scénarios évolutifs au sein de l'arbre modèle associe une grande quantité d'informations aux 427 gènes (214 actuels et 213 ancestraux) : leur composition en modules (1059 modules conservés) et leurs interactions possibles (471 partenaires

d'interactions protéine-protéine). La vaste quantité d'information que cela représente pose le problème de leur représentation afin de pouvoir les exploiter et les explorer. Le problème ici n'est pas l'espace disque utilisé par les fichiers, mais de faciliter l'interprétabilité et de permettre la visualisation des données produites. La Figure 8.3 présente un exemple de modèle factice que nous utiliserons pour illustrer les représentations.

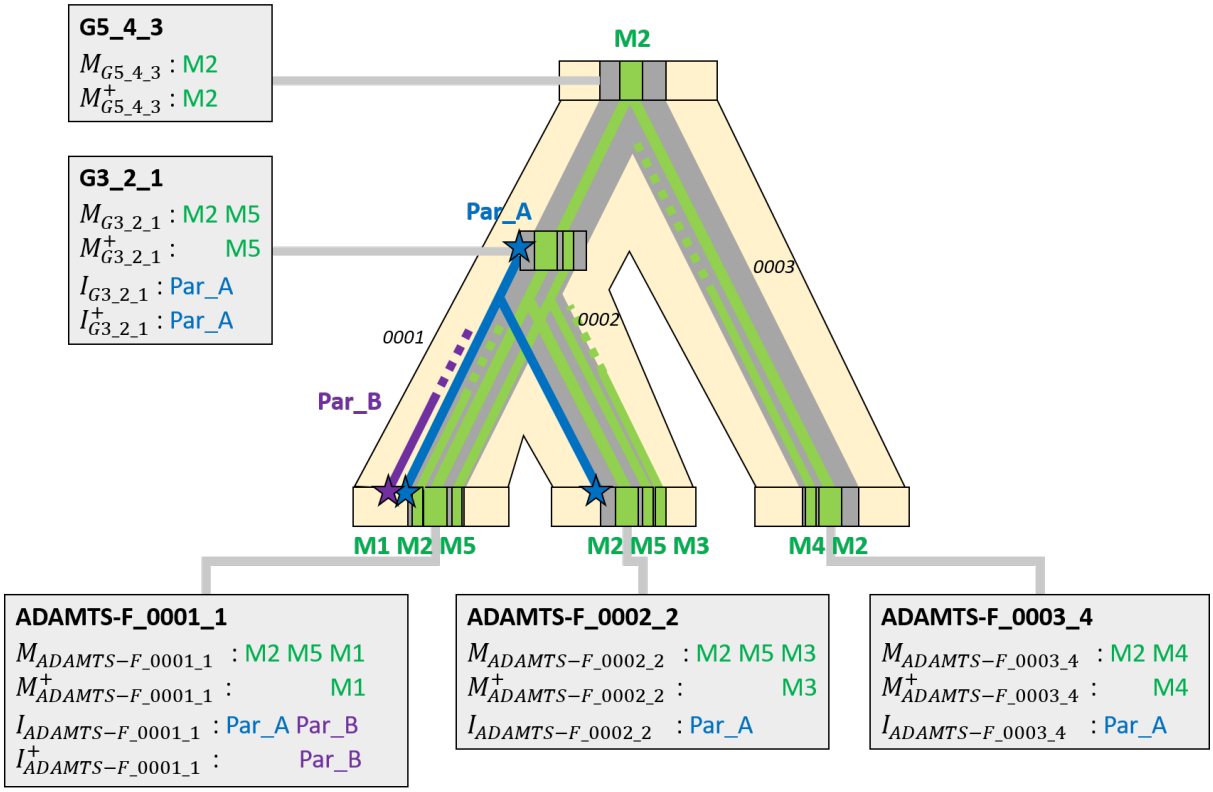

FIGURE 8.3 – **Exemple de modèle d'évolution jointe des modules et des interactions.** Exemple factice de modèle d'évolution jointe des modules (en vert) et des interactions (en bleu et violet) pour cinq gènes (en gris), deux ancestraux (G...) et trois actuels (ADAMTS-F...) appartenant à trois espèces (arbre en jaune, *0001, 0002, 0003*). Chaque gène-nœud n est décrit par la liste de ses modules présents $\mathcal{M}_n$, la liste de ses modules gagnés $\mathcal{M}^+_n$, la liste de ses modules perdus $\mathcal{M}^-_n$, la liste de ses interactions présentes $\mathcal{I}_n$, la liste de ses interactions gagnées $\mathcal{I}^+_n$ et la liste de ses interactions perdues $\mathcal{I}^-_n$. L'exemple factice présente cinq gènes, comprenant trois ancestraux et trois actuels, associés à cinq modules (M1, M2, M3, M4 et M5) et deux interactions (Par_A et Par_B). Les relations de parenté issues de la topologie de l'arbre des gènes sont stockées dans les noms des nœuds ; le nœud ancestral G**3**_2_1 est l'ancêtre des feuilles ADAMTS-F_0001_**1** et ADAMTS-F_0002_**2** (comme indiqué par le _2_1 dans son nom), et le nœud ancestral G**5**_4_3 est l'ancêtre de la feuille ADAMTS-F_0003_**4** et du nœud ancestral G**3**_2_1. Les formats de notations sont définis par l'implémentation de `SEADOG-MD`.

8.2.1 Description du modèle par des fichiers tabulaires

Notre approche la plus basique a consisté à représenter les données produites par le modèle sous leur forme la plus brute possible. Nous fournissons ces différentes informations sous forme de différents fichiers tabulaires, ce qui en permet une analyse détaillée. Dans un fichier tabulaire au format `csv`, le modèle est alors décrit de la manière suivante : un gène (ancestral ou actuel) correspond à une ligne et les différentes colonnes correspondent aux modules et interactions présents, gagnés et perdus, comme présentés dans l'exemple (Tableau 8.1). L'arbre des gènes de références au format `newick` permet alors une analyse complète de l'arbre et des relations entre les différents gènes, toutes les informations de parenté permettant d'en reconstruire la topologie sont contenues dans les noms des gènes, et donc dans le fichier tabulaire. En effet, suite à la réconciliation mDGS par `SEADOG-MD`, tous les gènes sont associés à des noms uniques dans lesquels leurs descendants sont renseignés. Par exemple, le gène-nœud ancestral G3_2_1 correspond au nœud 3 qui a pour descendant le nœud 1 et le nœud 2. Ce format des données permet le stockage et l'accès computationnel aux données, ce qui en fait la description la plus efficace pour automatiser une analyse. Il est par exemple possible de récupérer tous les nœuds auxquels une composition spécifique en module est présente, tous ceux où au moins un module est gagné et tous les modules qui sont spécifiquement gagnés en même temps qu'une interaction d'intérêt, etc.

TABLE 8.1 – **Représentation tabulaire du modèle.**
Exemple factice de représentation tabulaire des données issues du modèle. Chaque ligne correspond à un gène n (i.e., un nœud de l'arbre des gènes), actuel (ADAMTS-F_0000X_X) ou ancestral (GX_X_X).

n	$\mathcal{M}_n$	$\mathcal{M}_n^+$	$\mathcal{M}_n^-$	$\mathcal{I}_n$	$\mathcal{I}_n^+$	$\mathcal{I}_n^-$
ADAMTS-F_0001_1	M1|M2|M5	M1		Par_A|Par_B	Par_B	
ADAMTS-F_0002_2	M3|M2|M5	M3		Par_A		
G3_2_1	M2|M5	M5		Par_A	Par_A	
ADAMTS-F_0003_4	M4|M2	M4		Par_A		
G5_4_3	M2	M2				

8.2.2 Le problème de la visualisation des données

Un fichier tabulaire permet la description et la mise à disposition de la totalité des données. Mais il ne permet pas d'en extraire directement du sens ni de pouvoir l'interpréter facilement. Il est parfois plus efficace et intuitif d'explorer et de visualiser de manière interactive des données que d'essayer de raisonner sur des listes dans un tableau. Dans notre modèle, il faut pouvoir interpréter conjointement des informations de parenté entre gènes (l'arbre de référence), de conservation de séquences (les modules), de phénotypes (interactions protéine-protéine). Nous avons choisi de nous appuyer sur l'arbre de référence des gènes pour explorer et visualiser ces informations afin de permettre une meilleure compréhension de notre modèle.

8.2.3 Description du modèle via une visualisation interactive de l'arbre sur Itol

Afin de faciliter l'exploration des données de notre modèle, nous avons développé une visualisation interactive au format de la plateforme en ligne de visualisation d'arbre phylogénétique `Itol` [LB19a] (*Interactive Tree Of Life*). Nous mettons à disposition une visualisation interactive de l'arbre de référence des gènes ADAMTS-TSL (disponible ici : **Arbre Itol**), annoté avec toutes les informations associées (e.g., modules et interactions présents, gagnés, perdus). Explorer un arbre sur `Itol` est très intuitif et différentes pages d'aide de la documentation d'`Itol` en permettent un usage avancé. De plus, nous proposons également une documentation spécifique à notre visualisation et aux données que nous y avons incluses (cf. Annexe 10.2).

Dans un premier temps, toutes les données sont disponibles de manière interactive sur l'arbre des gènes qui sert alors de support, par le biais de fenêtres *popup* interactives (décrites en `html`) associées aux différents gènes-nœuds (Figure 8.4). La totalité des informations concernant l'évolution des gènes, des modules et des interactions y sont présentes sous la forme de listes : les interactions présentes, gagnées, perdues, les modules présents, gagnés et perdus. De plus, pour la protéine actuelle, ou pour les partenaires d'interactions, un lien interactif permet d'accéder aux entrées correspondantes sur le site internet du NCBI ou d'Uniprot.

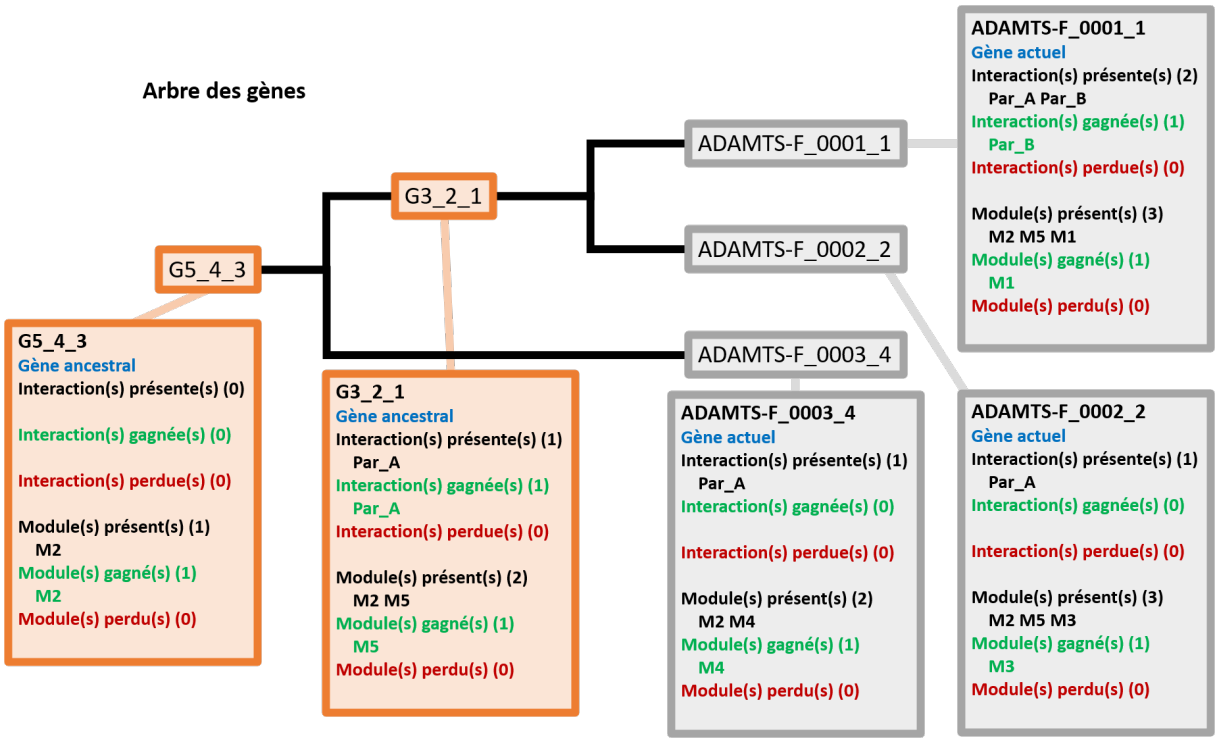

FIGURE 8.4 – **Représentation du modèle via l'arbre Itol.**
Exemple factice de représentation par arbre `Itol` des données du modèle. La plateforme en ligne de visualisation d'arbre phylogénétique `Itol` nous permet de visualiser l'arbre des gènes (en noir) qui comprend ici cinq gènes-nœuds, trois actuels (en gris) et deux ancestraux (en orange) avec leurs relations de parenté (branches de l'arbre). Chaque gène-nœud n est décrit avec un *popup* interactif qui présente les données des évolutions des modules et des interactions sous la forme des listes : interactions présentes $\mathcal{I}_n$, interactions gagnées $\mathcal{I}_n^+$, interactions perdues $\mathcal{I}_n^-$, modules présents $\mathcal{M}_n$, modules gagnés $\mathcal{M}_n^+$ et modules perdus $\mathcal{M}_n^-$.

Dans un second temps, nous utilisons la fonctionnalité de fichiers d'annotations d'`Itol` afin de construire des visualisations des différents types de données. En effet, `Itol` propose des formats propres à la plateforme qui permettent de visualiser les données qui y sont décrites. Par exemple, un fichier au format `Itol` d'annotation de domaines permet de décrire des domaines/régions/motifs à partir de leurs positions, formes et couleurs et ainsi de les visualiser sur un arbre `Itol`. Au moment de la rédaction de cette thèse (octobre 2022), il existait 17 types de formats d'annotation différents [1]. Afin de permettre une visualisation de nos données de gènes, de modules, d'interactions et leurs évolutions, nous les avons alors décrites dans les formats d'annotation `Itol` adaptés (Figure 8.5). Ces fichiers d'annotations nous permettent de générer un grand nombre de visualisations et de figures (qui sont présentées dans ce manuscrit).

En conclusion, nous proposons deux types de représentation (tabulaire et interactive avec `Itol`) des données du modèle d'évolution jointe des modules et des interactions. Ces représentations sont toutes deux utiles et complémentaires, tout particulièrement quand elles sont utilisées en simultané, et rendent alors possible d'explorer, comprendre, d'extraire de manière automatique avec des scripts et de visualiser les informations d'intérêts, tout cela avec la possibilité de construire une figure rapidement grâce à `Itol`. À noter que tous les fichiers qui décrivent ces deux représentations (`csv`, `newick`, fichiers `txt` au format d'annotation `Itol`) sont générés automatiquement à la fin du *framework* `PhyloCharMod` (voir Section 8.1.1).

1. Décrits à l'adresse suivante : https://itol.embl.de/help.cgi

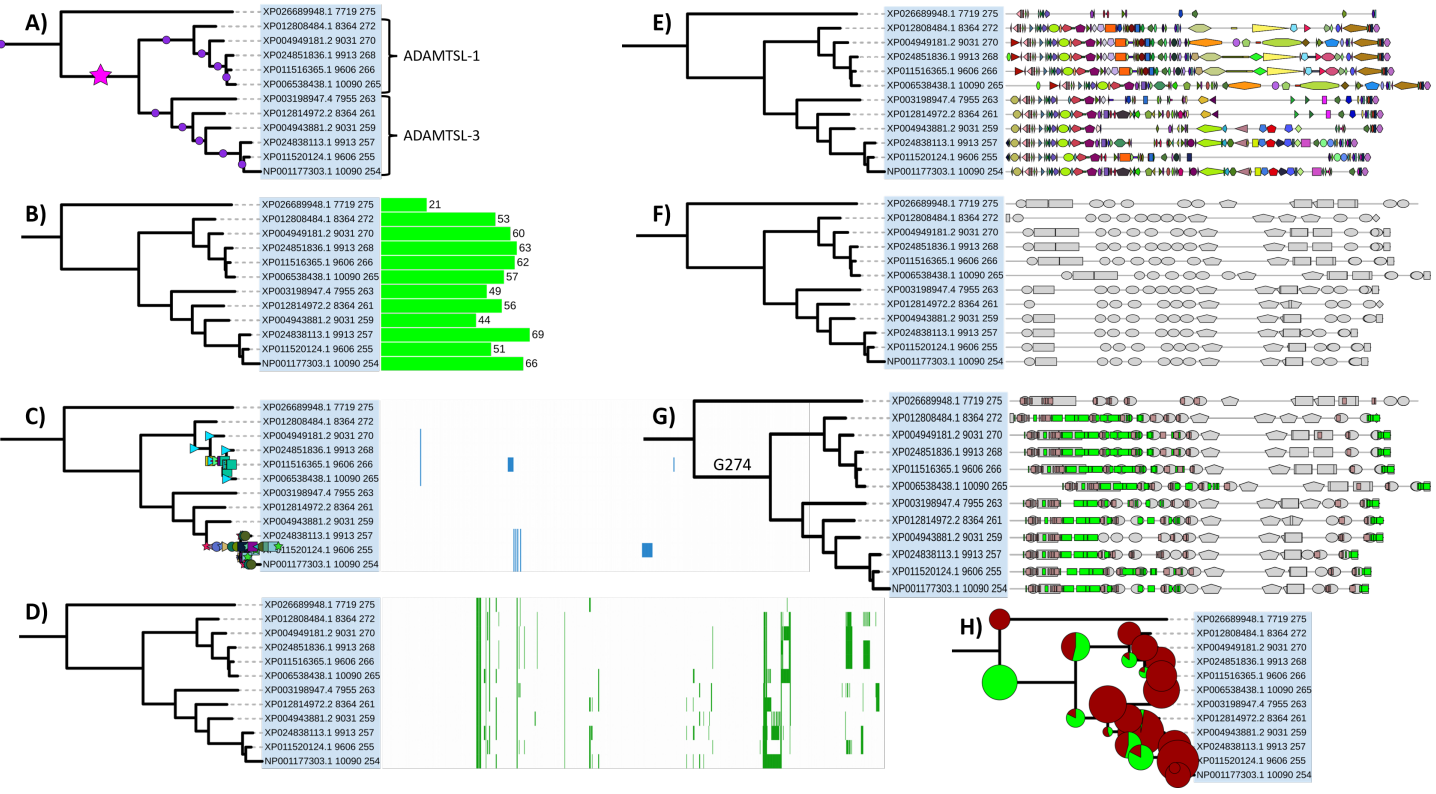

FIGURE 8.5 – **Visualisation des données du modèle via les annotations Itol.**
Huit exemples de visualisation des données du modèle aux formats d'annotations d'`Itol`, représentés sur le sous-arbre des gènes homologues de ADAMTSL-1 et ADAMTSL-3 (sous-ensemble de l'arbre des gènes de référence). **A)** Événements de spéciations (rond) et duplications de gènes (étoile), d'après les *mapping Gène-Espèce* de la réconciliation Module-Gène-Espèce. **B)** Nombre de modules par gène actuel, d'après la segmentation `Paloma-2`. **C)** Matrice de la présence des interactions à chaque gène actuel (en bleu, alignée aux feuilles) et présence des interactions à chaque gène ancestral sur l'arbre, basé sur les scénarios ancestraux `PastML`. **D)** Matrice de la présence des modules à chaque gène actuel (en vert, alignée aux feuilles). **E)** Composition en modules des gènes actuels. **F)** Composition en domaines des gènes actuels. **G)** Modules présents (marron) et gagnés (vert) au gène ancestral G274, avec la composition en domaines en fond (gris). **H)** Quantification du nombre de modules gagnés/perdus à chaque gène ancestral.

8.3 Coapparition Module-Interaction

Un module est une région fortement conservée, partagée entre plusieurs protéines et que nous supposons résulter d'une forte pression de sélection liée au maintien d'une fonction de la protéine. De plus, la divergence d'une région fonctionnelle est corrélée à la divergence de la fonction dans laquelle elle est impliquée [Gau+11 ; FG13 ; Lau+21]. Quand un module diverge, nous supposerons que sa fonction diverge également. Notre modèle joint les évolutions des modules et les évolutions des interactions, ce qui nous permet d'étudier les divergences modules-interactions concomitantes, afin d'associer modules et fonctions. De la même manière dont nous avons regroupé des modules sur la base leurs acquisitions simultanées (signature de modules) au cours de l'évolution des gènes, nous avons associé les modules et interactions gagnés à un même moment de l'évolution des gènes afin d'identifier des événements de *coapparitions module(s) - interaction(s)*. Nous définissons un événement de coapparition comme un nœud ancestral de l'arbre des gènes (i.e., gène ancestral), où au moins un module et une interaction sont gagnés (Figure 8.6). Notre hypothèse est alors que les modules et les interactions gagnés à un même moment de l'évolution d'un gène sont liés. Ces modules pourraient être impliqués dans les interactions avec lesquelles ils co-apparaissent : quand une région fonctionnelle est acquise et conservée au cours de l'évolution, sa fonction est acquise et conservée de manière concomitante.

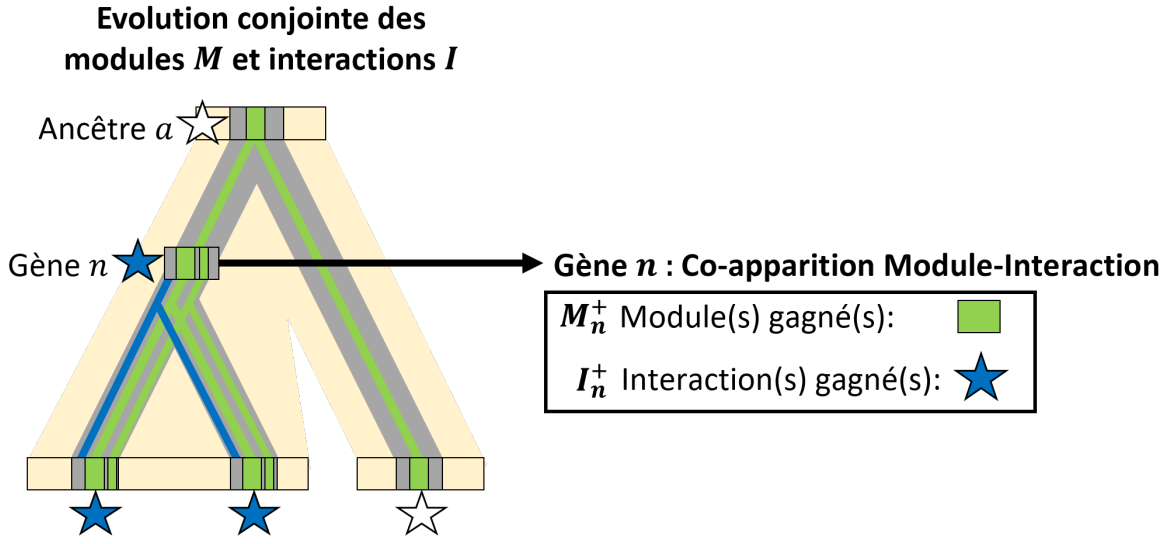

FIGURE 8.6 – **Événement de coapparition module(s)-Phénotype(s).**
Un gène pour lequel au moins un module et au moins une interaction sont gagnés est considéré comme un gène pour lequel un événement de coapparition module-interaction s'est produit.

⇒ Les résultats détaillés dans la fin de ce chapitre sont également présentés dans l'article :
Phylogenetic prediction of functional modules in ADAMTS-TSL proteins reveals sequence signatures involved in protein-protein interactions.
O Dennler, F Coste, S Blanquart, C Belleannée, N Théret. (2022).
Submitted in PLOS Computational Biology (19/09/2022)

8.3.1 Identification de 45 événements de coapparition modules-interactions

Parmi les 213 gènes ancestraux de l'arbre de référence ADAMTS-TSL, nous avons identifié 183 ancêtres où au moins un module est gagné (voir Section 6.4.3), et 48 où au moins une interaction est gagnée (voir Section 7.3.2). Nous identifions alors 45 gènes ancestraux correspondant à un événement de coapparition module(s) - interaction(s). À noter que seulement trois gènes ancestraux présentent une acquisition d'interactions sans gain de modules (G149, G217, G329). Nous distinguons deux types de gènes ancestraux présentant un évènement de coapparition : des gènes ancêtres de plusieurs copies paralogues (on en décompte cinq, Tableau 8.2), et des gènes ancêtres de protéines orthologues (on en décompte quarante, Tableau 8.3). Ces 45 événements de coapparition impliquent au total 355 modules et 278 interactions, variant d'un événement à l'autre de 1 à 26 modules, et de 1 à 143 interactions. Certaines interactions sont acquises plusieurs fois au cours de l'évolution des ADAMTS-TSL, suggérant la possibilité d'événements de convergences évolutives (i.e., homoplasie).

TABLE 8.2 – **Les 5 événements de coapparition de module(s)-interactions(s) correspondant à des gènes ancestraux de plusieurs copies paralogues.**
Noms des gènes tels que sur la Figure 8.7.

Gène	Descendant(s)	Interactions(s)	Module(s)
G91	ADAMTS-9, -20	LRP1	B1715 B858 B949 B1958 B888 B946 B1707 B1283 B854 B836 B987 B835 B911 B860 B856 B889 B1704 B398 B913 B820 B1705 B1728 B939 B865 B910 B400
G96	ADAMTS-1, -4, -5, -8, -9, -15, -20	VCAN ACAN	B773 B781
G161	ADAMTS-7, -12	GRN CCN2 COMP	B1289 B1456 B1331 B1272 B1460 B1993 B594 B592 B1305 B1429 B1283 B1332 B911 B1269 B1317 B1306 B1811 B1315 B1307 B1995 B1288 B1328 B1316 B1247 B910
G235	ADAMTSL-4, -6	NUFIP2	B2173 B1188 B2168 B2153 B592 B2174 B2203 B2139 B2210 B2147 B2199 B2200 B2201 B2172
G341	ADAMTS-2, -3, -14	COL5A1 COL1A2 TLL1 F13A1 CFB C3 KNG1 IGHG3 COL1A1 COL2A1 COL3A1 COL4A1 APOA1 APOA2 APOH FN1 AHSG GC ANXA1 COL5A2 KLK1 ANXA2 MMP3 CFH VIM LGALS1 CTSH C4B COL11A1 COL6A1 COL6A2 BMP1 ANXA8 COL11A2 MIF TCN2 OGN BGN MPZ CTSS COL5A3 GRN HLAA SERPINB6 LGALS7 LUM ACTA2 HBB CAV1 TGFBR3 COL14A1 DPT TMED1 ITIH4 POSTN PDIA6 PCOLCE CDSN COL24A1 SUSD4 DMKN CAVIN1 AEBP2 COL27A1 UNC80 CPA4	B1404 B1340 B1378 B1350 B1391 B1360 B1380 B1492 B1407 B1415 B1399 B1341 B1394 B1362 B1349 B1393 B1403 B1374 B1483 B1390 B1637 B1409

TABLE 8.3 – **Les 40 événements de coapparition de module(s)-interactions(s) correspondant à des gènes ancestraux d'une seule copie paralogue.**
Noms des gènes tels que sur la Figure 8.7.

Gène	Descendant(s)	Interactions(s)	Module(s)
G4	ADAMTS-1	HPX DKK3	B561 B1646 B1661 B1657 B1659 B608 B1658 B605 B1668 B1647 B1645
G8	ADAMTS-1	A2M	B1673
G10	ADAMTS-1	FURIN	B687 B882 B1283 B1665 B1672 B686
G13	ADAMTS-4	RELN	B1950
G15	ADAMTS-4	A2M FN1 BGN COMP LRP1	B1693 B1675 B1678 B1497 B1674 B1496 B1692 B1676
G19	ADAMTS-4	FURIN	B1691 B1689 B1677 B1687 B1842 B1683
G23	ADAMTS-8	SPP1	B1545 B1552
G59	ADAMTS-5	NCAN CILP MATN4 FMOD SDC1 ITIH2 TNC TIMP3 MMP13 FBLN2 CILP2 ADAMTS5	B1583 B1595 B1610 B1945 B1575 B1577 B1598 B1944 B1591 B1592
G61	ADAMTS-5	DCN PRELP	B491
G63	ADAMTS-5	COL2A1 COL3A1 FN1 BGN RELN	B657 B1578 B658 B1608
G65	ADAMTS-5	A2M LRP1	B1602 B1605 B1593 B1594 B2029 B1588 B1601 B1607 B1617 B1603 B1580 B686
G71	ADAMTS-9	RNF123	B522
G75	ADAMTS-9	FN1	B974 B1734 B1714 B976 B973 B963 B983 B917 B993 B978 B1702 B873 B1708
G103	ADAMTS-6	ERP29 SDC4 NEK4 SNTA1 LTBP1 RNF2	B1864 B919
G114	ADAMTS-10	FBN1	B671 B676 B683 B606 B673 B702 B674
G128	Papilin	PKM THOP1 DPY19L3 KLHL36 ZNF507 TFAP2C KIF2C NIF3L1	B2223 B2233 B2265 B2253 B2222 B2214 B2268 B2257 B2241 B2240 B2255 B2251 B2228 B2254 B2232 B2252
G141	ADAMTS-12	HELS72 FKBP9 CALM2 USF1 PPM1A MEOX2 MYL6 SP3 USF2 RUFY3 UBR1 SNRK	B1318
G143	ADAMTS-12	NCAN	B1298 B1262 B687 B1249 B1320 B1296 B1287 B1290 B1286 B1299 B1253 B1300 B1260 B1291 B1325 B1336 B1245 B1297
G152	ADAMTS-7	CTSB	B560 B1919 B1918 B1917
G154	ADAMTS-7	COL5A1 COL1A2 CST3 COL1A1 COL3A1 FN1 COL5A2 COL6A3 HLAA	B1450 B1454 B1444 B1436 B1287 B1426 B1470 B1290 B1286 B1453 B1455 B1443 B1433 B1451 B1435 B1430
G160	ADAMTS-7	A2M	B1923 B1431 B1921 B1924 B1432
G169	ADAMTS-18	B3GLCT	B1096
G171	ADAMTS-18	ALB	B477 B1073 B1095
G182	ADAMTS-16	FN1	B1025 B1012 B1057 B1893 B1004 B1003 B1894
G221	ADAMTSL-6	FBN1	B2130 B2163 B2155 B2145 B519 B2353 B2160 B2156 B2180 B2132 B2140 B509 B2124
G228	ADAMTSL-4	GLRX3 KRTAP108 KRTAP11	B2419 B2422 B2417 B2418 B2421 B2420
G230	ADAMTSL-4	EFEMP2 COL5A1 ITGB4 EGFL9 COL1A2 CLEC18A CYSRT1 US26 LCE2B PLSCR1 ADAM12 FRS3 SPRY2 PRKAB2 TLL1 MAPKBP1 SORBS3 EIF4E2 STK16 FARS2 BAG4 TOP3B COL1A1 COL2A1 COL3A1 ITGB2 COL5A2 CTSB CST2 GIP COL11A1 BMP1 COL11A2 HGF FAH VCAM1 SPINK2 FLNA GATA2 LMO2 LMO1 COL5A3 KRTAP59 COL8A1 CFP AQP1 HOXC8 OTX1 NTF4 PTGER3 HOXA1 FXR1 KRTAP101 KRTAP103 KRTAP105 KRTAP109 KRTAP1011 FKBP1B CREB5 TNK2 KIF1A PIN1 FHL3 DGCR6 DIP2A GNMT MVP TCEA2 TRIP6 LONRF1 COL24A1 KRTAP212 MGAT5B MIIP LCE3C LCE3E LCE1C LCE1B LCE4A LCE2D VASN KRTAP56 MORN3 LRFN4 SPATA8 DLK2 ADAMTSL4 ADAMTSL5 KRTAP52 KCTD9 PID1 ATG9A NEK8 TMEM150A TRIM42 CFAP206 COL27A1 NATD1 CATSPER1 DBF4B ZNF417 LGALS14 RAB2B LRRC29 MYLIP SUSD6 SMARCC1 FBXW5 DISP1 NTAQ1 OLFM3 RCHY1 TSSK3 ZNF587 CYP2S1 KRTAP412 APOL6 TAPBPL DGCR6L KRTAP94 KRTAP92 KRTAP411 KRTAP42 ASPSCR1 NMUR2 CXCL16 RHOJ BANF2 FAM124B ARNT2 SLC6A20 FBXO6 SHFL CPNE7 TUBGCP4 SLC23A1 MID2 DNPEP GNE EXOSC1 SALL2 AMMECR1 CHCHD2	B2182 B2181 B2183 B2195 B2188 B2194 B2189 B2196 B2208
G248	ADAMTSL-2	FBN1 HOXA1 LOXL3 NECAB2	B2310 B2378 B2377 B2299 B2385 B2301 B2379 B2380 B1820 B2381
G258	ADAMTSL-3	CYSRT1 GLRX3 KRTAP23 NOTCH2NLC KRTAP123 KRTAP103 KRTAP106 KRTAP108 KRTAP11 KRT40 KRTAP57 NOTCH2NLA MDFI KRTAP32 KRTAP24	B2084 B2071 B657 B2091 B2076 B2082 B2074 B519 B2102 B1287 B2079 B2042 B2391 B2072 B2395 B2394 B2078 B2087 B2393
G269	ADAMTSL-1	MMP10 FHL2 ACOX1 B3GLCT RSPRY1 WDCP	B2401 B2407 B2403 B2405 B2102 B2411 B2096 B2409
G271	ADAMTSL-1	GRN	B2105 B2114 B2098 B2107 B2113 B2110 B2106 B2109 B2112 B2104 B2108 B2344
G281	ADAMTSL-5	CYSRT1 KRAS NOTCH2NLC KRTAP59 FBN1 KRTAP103 FHL5 ADAMTSL4 NOTCH2NLA	B2283 B2285 B2280 B2275 B2290 B2288 B2434
G299	ADAMTS-13	F8 VWF	B1191 B1180 B1206 B1159 B1190 B1204 B1181 B1197 B1199 B1900 B1897
G315	ADAMTS-3	CCN2	B1424
G324	ADAMTS-2	ALB HPX DCN CTSB DKK3	B1621 B1640 B1623 B1618 B1627 B1620 B1629
G326	ADAMTS-2	A2M	B1641 B1638 B510 B1626 B1643
G328	ADAMTS-2	CST3 COL6A3 PRELP	B1644 B563 B1636 B1631 B1632 B1835
G334	ADAMTS-14	A2M ALB HPX DCN CTSB VCAN DKK3	B1485 B1499 B1486 B1497 B1493 B1410 B1496 B1481 B1498 B1472
G336	ADAMTS-14	CST3 COL6A3	B1480
G340	ADAMTS-14	PRELP	B1474

8.3.1.1 Diversité des événements de coapparitions

Les associations module(s) / interaction(s) obtenues sont très diverses, certaines associent une faible quantité de modules à une faible quantité d'interactions. C'est par exemple le cas du gène ancestral G96 qui associe le gain de deux interactions (avec ACAN et VCAN) au gain de deux modules (B773, B783), ou le cas du gène ancestral G315 qui associe le gain de la seule interaction CCN2 au gain du seul module B1424. À l'opposé, le gène ancestral G341 associe le gain de 66 interactions au gain de 22 modules. Nous observons également des associations plus déséquilibrées, par exemple celle du gène ancestral G91 qui associe 26 modules à l'interaction avec la protéine LRP1. Ces diversités traduisent des espaces de recherche très variables dans les associations modules-interactions qui résultent de notre modèle. En effet, associer un unique module à une unique interaction permet de proposer un motif candidat très précis. Proposer plusieurs modules pour une unique interaction permet de proposer une interface plus complexe comme candidats de l'interaction. Il est plus compliqué cependant d'interpréter une association d'un grand nombre de modules à un grand nombre d'interactions. Ce qui pose les questions suivantes : tous les modules sont ils impliqués dans toutes ces interactions ? Différents modules sont ils impliqués dans les différentes interactions ?

8.3.1.2 Localisation des modules impliqués dans des événements de coapparition

Nous nous sommes ensuite intéressés aux modules impliqués dans les 45 événements de coapparition (Figure 8.7). De manière à se focaliser sur l'humain, les modules gagnés ancestralement sont représentés par les segments de leurs descendants humains. Parmi les 355 modules impliqués dans les 45 coapparitions, 283 sont présents dans les descendants humains sélectionnés, les modules non-présents peuvent avoir été perdus ou avoir divergés subséquemment dans la lignée humaine. Nous observons que les positions des modules gagnés ne correspondent pas nécessairement aux positions des domaines Pfam ; les modules associés aux interactions étant localisés à l'intérieur, en chevauchement, ou à l'extérieur des domaines. Cependant, nous montrons que la « région ancillaire variable » (colorée en violet sur la Figure 8.7) contient le plus grand nombre de modules gagnés (49 %). Cette observation est consistante avec la variabilité de la région ancillaire et son implication connue dans la spécificité des interactions chez les ADAMTS-TSL [Kel+15]. Nous montrons également que 30 % des modules gagnés sont localisés au sein de la région N-terminale (colorée en bleu sur la Figure 8.7), qui correspond principalement au propetide et aux résidus qui l'entourent. Il est intéressant de noter que le propeptide est connu pour évoluer plus rapidement que les domaines de manière à permettre une spécialisation des fonctions de la protéine [Dem+10]. À l'opposé, la région centrale, qui inclue les domaines métalloprotéase et disintegrine des ADAMTS (colorée en orange et en jaune sur la Figure 8.7) présente une proportion bien moins importante de modules gagnés (20 %, 58 parmi 283). La majorité des modules de la région centrale est concentrée au sein du domaine spacer des ADAMTS-TSL, qui contient alors 26 des 58 modules gagnés dans la région centrale. Finalement, nos données mettent en évidence des gains de modules potentiellement associés aux interactions spécifiquement au sein des régions N-terminale et C-terminale et du domaine spacer des ADAMTS-TSL.

Tous ces évènements de coapparition se produisent à des nœuds récents (proche des feuilles), les plus anciens correspondant à des ancêtres de sous-groupes fonctionnels (G96 pour l'ancêtre des hyalectanases et G341 pour l'ancêtre des procollagénases). Tous ces nœuds de coapparitions correspondent ainsi à des bipartitions que nous avons identifiées comme extrêmement robustes au rééchantillonnage taxonomique, à la méthode d'inférence et à la variation du groupe externe. Nous avons étudié en détails différents événements de coapparition, en particulier ceux impliquant plusieurs copies paralogues. Notre objectif principal était ici de confronter nos prédictions de modules fonctionnels avec la bibliogra-

phie disponible afin de valider notre modèle.

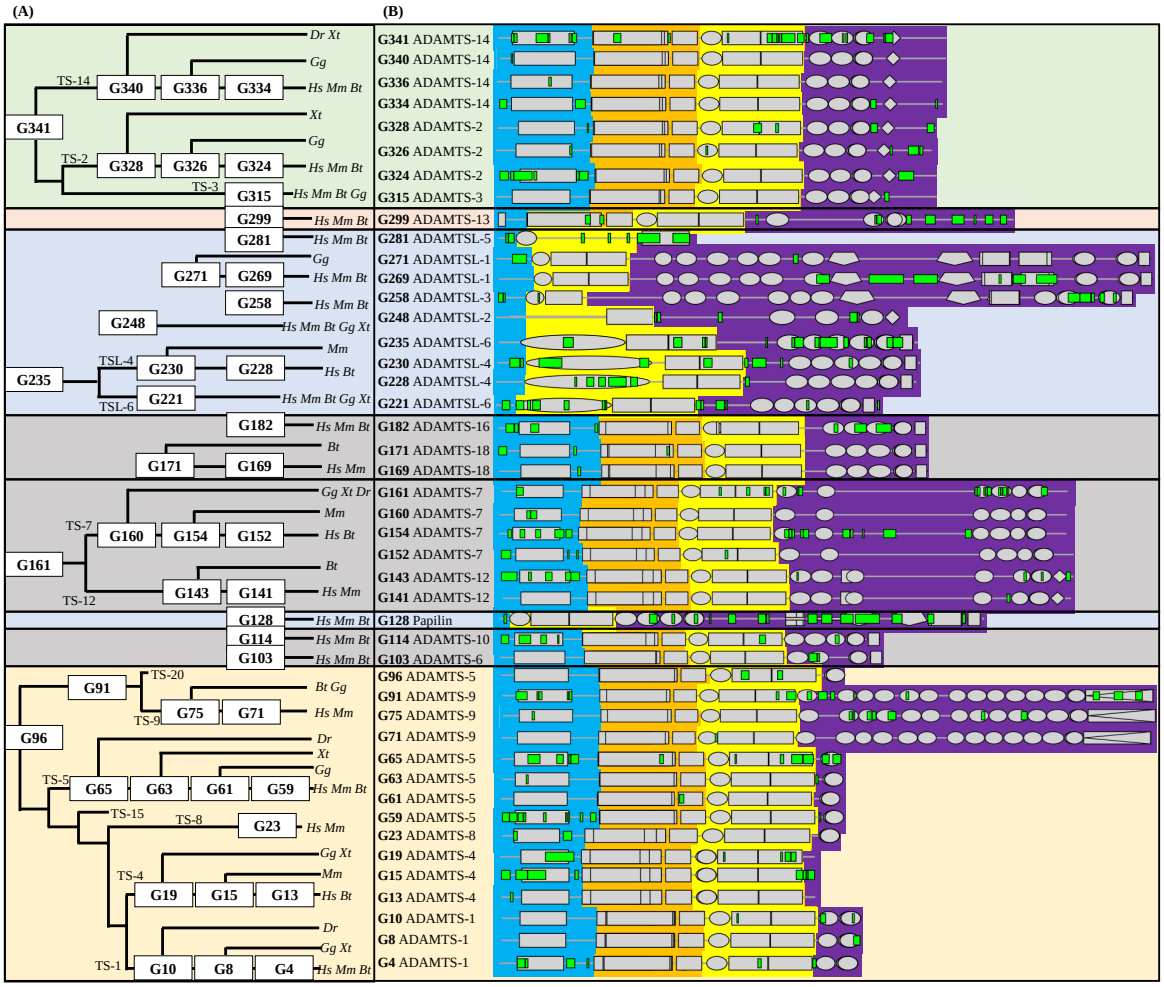

FIGURE 8.7 – **Localisation des modules impliqués dans les 45 événements de coapparition module(s)-interaction(s), sur les descendants humains.**
(A) Phylogénie simplifiée des ADAMTS-TSL représentant les 45 gènes ancestraux (boîtes blanches) qui correspondent aux 45 événements de coapparition module(s)-interaction(s). TS : ADAMTS ; TSL : ADAMTSL ; *Hs, Homo sapiens* ; *Dr, Danio rerio* ; *Xt, Xenopus tropicalis* ; *Gg, Gallus gallus* ; *Mm, Mus musculus* ; *Bt, Bos taurus*. (B) Chaque ligne correspond à une protéine humaine choisie comme représentante d'un des 45 ancêtres étudiés, sur laquelle sont reportés les segments actuels des modules gagnés au gène ancestral correspondant. Leurs compositions en domaines sont représentées en formes grises et les modules gagnés en rectangles verts. Chaque séquence est divisée en trois régions ; 1) la région N-terminale qui contient le propeptide (**bleu**), 2) la région centrale qui inclue le domaine catalytique et le domaine disintegrine (**orange**) ainsi que le TSP1 central, le domaine cystéine riche et le spacer (**jaune**), et 3) la région ancillaire variable, allant de la fin du spacer à l'extrémité C-terminale (**violet**).

8.3.2 Convergence évolutive des interactions avec les protéines COMP et CCN2

Parmi les 26 ADAMTS-TSL humaines, nous avons identifié ADAMTS-3, ADAMTS-4, ADAMTS-7 et ADAMTS-12 comme capables d'interagir avec la protéine COMP (*Cartilage Oligometric Matrix Protein*) et/ou avec CCN2 (alias CTGF, *Connective Tissue Growth Factor*). Le rôle d'ADAMTS-7 et ADAMTS-12 dans la pathogenèse de l'arthrite a dans un premier temps été décrit par leur capacité commune à cliver COMP [Liu+06a ; Liu+06b]. Plus récemment, la protéine CCN2 a également été identifiée comme substrat d'ADAMTS-7 et ADAMTS-12 [Pi+15 ; Wei+18]. Les compositions en domaines des protéines ADAMTS-7 et ADAMTS-12 étant très similaires, elles forment leur propre sous-groupe, il n'est ainsi pas surprenant qu'elles partagent des substrats [Liu09]. Par contre, il est remarquable d'identifier CCN2 et COMP comme substrats d'ADAMTS distantes évolutivement, à savoir ADAMTS-3 [Bek+16] et ADAMTS-4 [Dic+03], ce qui suggère des évènements convergent d'apparitions des interactions avec COMP et CCN2.

8.3.2.1 Acquisitions multiples des interactions avec COMP et CCN2

Comme présenté en Figure 8.8 et selon nos inférences `PastML`, les interactions avec CCN2 et COMP seraient apparues deux fois chacune au cours de l'évolution des ADAMTS-TSL. D'une part, ces deux interactions ont toutes les deux été acquises au gène ancestral G161, l'ancêtre commun des gènes ADAMTS-7 et ADAMTS-12. L'interaction avec COMP a également été acquise au gène ancestral G15, l'ancêtre des orthologues *H. sapiens, B. taurus, M. musculus* de ADAMTS-4. D'une manière similaire, l'interaction avec CCN2 a été acquise au gène G315, l'ancêtre des orthologues *H. sapiens, B. taurus, M. musculus, G. gallus* de ADAMTS-3.

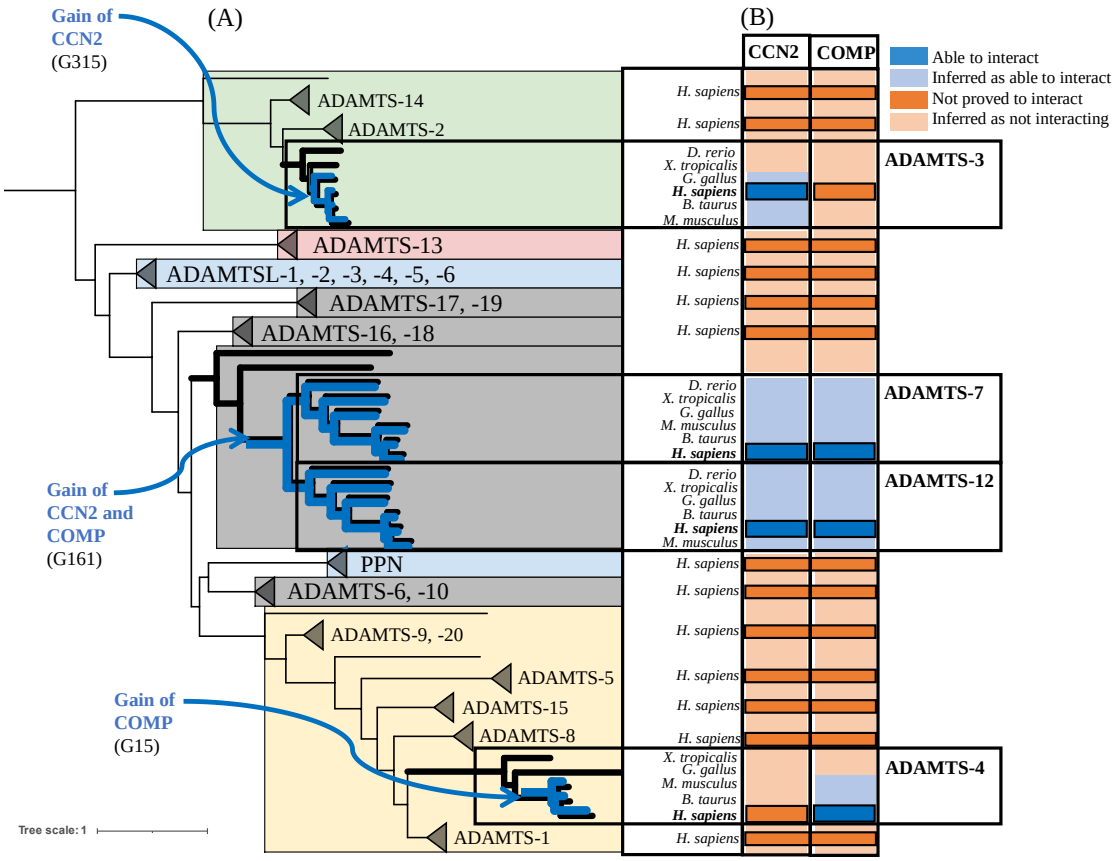

FIGURE 8.8 – **Convergence évolutive des interactions ADAMTS-COMP et ADAMTS-CCN2.**
(A) Arbre phylogénétique de référence des ADAMTS-TSL, détaillé pour les sous-arbres impliqués dans les interactions COMP et CCN2. (B) Représentation des interactions avec COMP et CCN2 issues de la littérature (bleu foncé). Les gains des interactions (inféré avec `PastML`) sont représentés par les nœuds internes : G315 (ancêtre *Amniota* des ADAMTS-3), G161 (ancêtre des paralogues ADAMTS-7 et ADAMTS-12) et G15 (l'ancêtre *Mammalia* des ADAMTS-4), correspondant respectivement aux interactions CCN2, CCN2/COMP et COMP. Les interactions inférées aux feuilles (protéines non humaines) sont représentées en bleu clair. À l'opposé, l'absence d'une interaction est représentée en orange foncé, et l'absence inférée par `PastML` en orange clair.

8.3.2.2 Trois signatures en modules distinctes

Les modules qui co-apparaissent avec les interactions COMP et CCN2 sont les modules gagnés aux trois gènes ancestraux : G161, G15, G315. Pour ce qui du gène ancestral G161, 25 modules sont gagnés, principalement dans les répétitions TSP1 de la région ancillaire (Figure 8.9 A). Les quatre répétitions TSP1 de cette région correspondent à une région ayant été identifiée comme suffisante et nécessaire aux interactions de ADAMTS-7 et ADAMTS-12 avec COMP [Liu+06a ; Liu+06b] et CCN2 [Pi+15 ; Wei+18]. En revanche, COMP a également été identifiée comme un substrat d'ADAMTS-4, qui ne possède pas ces quatre répétitions TSP1 dans la région C-terminale, ce qui suggère que l'interaction COMP est associée à des séquences/motifs différents. En accord avec cette hypothèse, nous avons identifié huit modules gagnés au gène ancestral G15 (Figure 8.9 B), qui co-apparaissent également avec l'interaction COMP. Ces huit modules sont localisés à l'extrémité N-terminale (incluant une partie du propeptide) ainsi qu'en C-terminale de la séquence d'ADAMTS-4. Aucun des modules gagnés à G15 n'est gagné à G161. D'une manière similaire, un module unique est gagné au gène ancestral G315 (Figure 8.9 C). Ce module se localise à l'extrémité C-terminale et n'est pas non plus gagné à G161. Ces analyses démontrent que, bien que majoritairement situées dans les régions ancillaires de leurs ADAMTS respectives, les trois signatures de modules (G161, G15, G315) diffèrent fortement par la composition en modules, les séquences des segments de ces modules, le contexte des domaines où ils se localisent et le contexte tridimensionnel prédit dans lequel ils se localisent (Figures 8.9 et 8.10).

Finalement, nos données supportent l'hypothèse d'acquisitions indépendantes des interactions COMP et CCN2, acquises chacune à deux moments distincts de l'évolution des ADAMTS-TSL. Dans le cadre de ces deux cas d'homoplasie, nous proposons trois signatures de modules fonctionnels, toutes caractérisées par des modules spécifiques dans l'extrémité C-terminale de la région ancillaire des protéines humaines ADAMTS-TSL. Ainsi, pour le cas COMP/CCN2, les acquisitions indépendantes sont toutes associées à de nouveaux sites d'interactions ADAMTS-TSL.

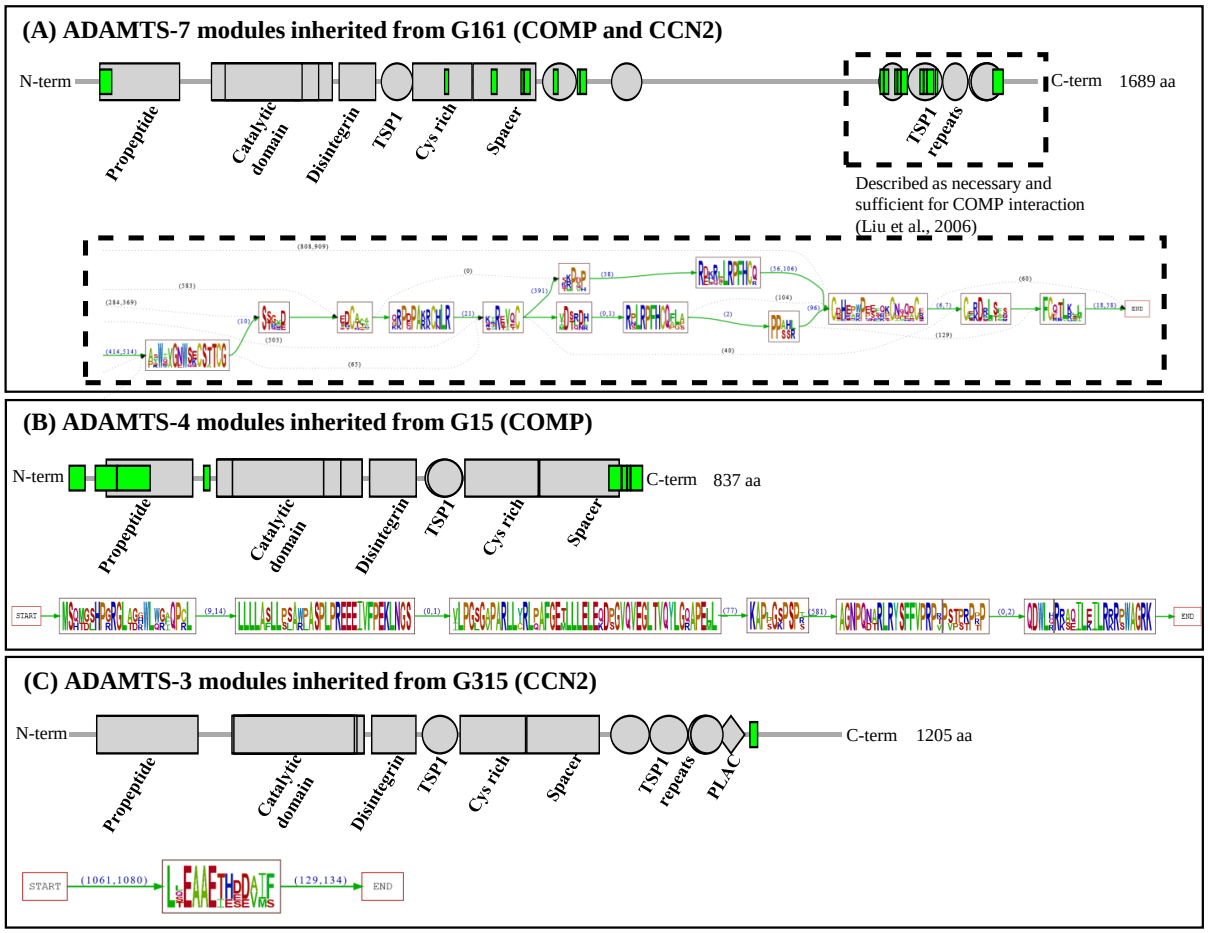

FIGURE 8.9 – **Trois signatures en modules distinctes sont associées aux interactions avec COMP et/ou CCN2.**
Les segments actuels des modules gagnés aux trois gènes ancestraux G161, G15 et G315 sont visualisés sur leurs protéines descendantes humaines : (A) sur ADAMTS-7, (B) sur ADAMTS-4 et (C) sur ADAMTS-3. Les segments de modules sont représentés en vert, les domaines en gris (partie supérieure). Les motifs d'acides aminés correspondants aux différents segments de chaque module sont représentés sous forme de logos (partie inférieure). L'intervalle en acides aminés séparant deux modules gagnés chez les différents descendants d'un même gène ancestral est représenté par une flèche, avec la taille de l'intervalle indiquée en bleu.

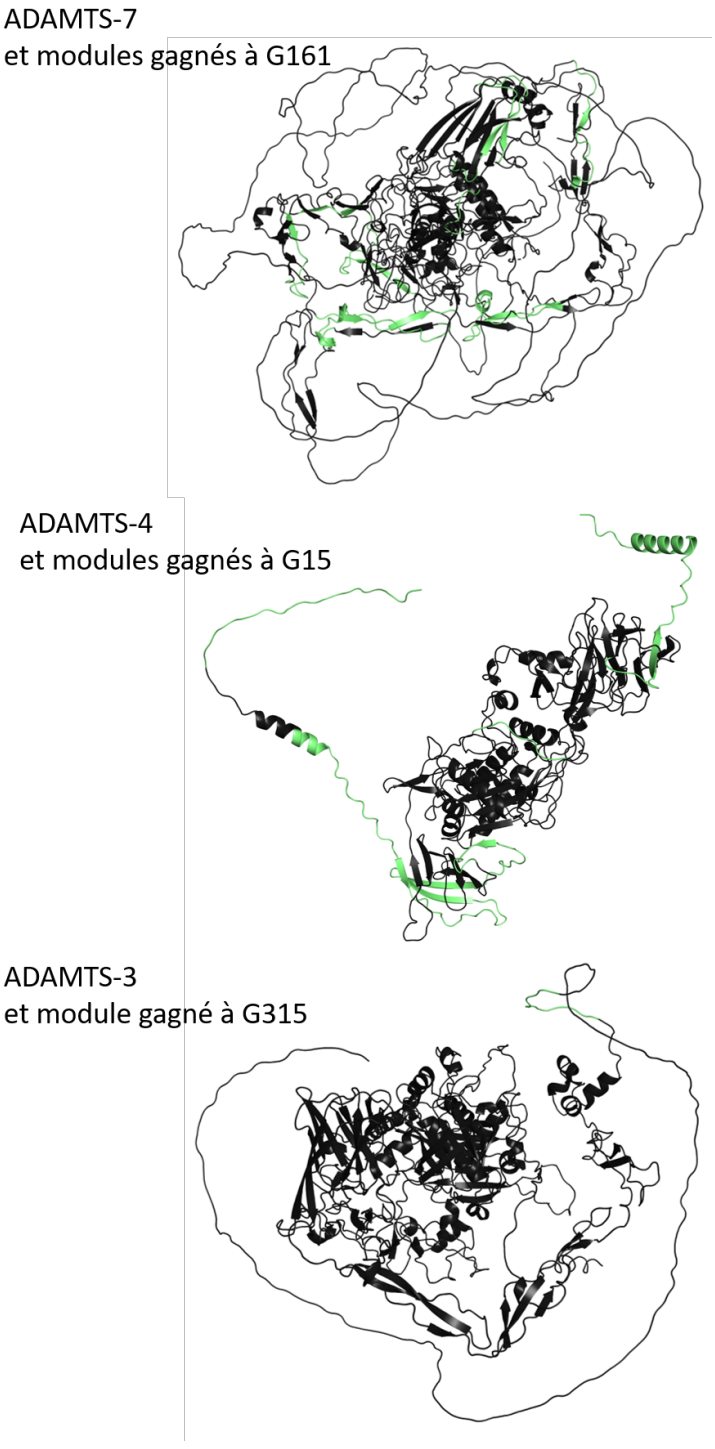

FIGURE 8.10 – **Projections sur les structures prédites des modules gagnés à G161, G15 et G315.**
Les modules gagnés (en verts) à G161, G15 et G315 sont projetés sur les structures prédites par `AlphaFold` de ADAMTS-7, ADAMTS-4 et ADAMTS-3 respectivement. Les modules gagnés seraient exposés (à l'extérieur de la protéine) mais se localisent dans des contextes tridimensionnels non comparables, ces trois structures prédites ne se ressemblant pas.

8.3.2.3 La réinvention d'interactions comme acteur de la complexité de la matrice extracellulaire ?

Les interactions protéine-protéine jouent un rôle important dans la structure de la matrice extracellulaire [Cro+14 ; Kar+21], c'est pourquoi il est très intéressant d'observer ces cas d'acquisitions indépendantes des interactions ADAMTS avec COMP et CCN2 qui impliqueraient des sites d'interactions totalement différents chez des protéines ADAMTS. En effet, l'interaction COMP implique les quatre motifs TSP1 répétés de ADAMTS-7 et ADAMTS-12 [Liu+06a ; Liu+06b], motifs qui sont absents chez ADAMTS-4, cette observation exclue totalement la possibilité d'une interaction ADAMTS-4/COMP qui utiliserait un site d'interaction similaire à celui de ADAMTS-7 et ADAMTS-12. Cette observation renforce l'hypothèse d'interactions ADAMTS/COMP impliquant des sites complètement différents. Nous pouvons supposer que des sites d'interactions différents permettent des fonctions différentes pour les couples d'interactions ADAMTS-4/COMP et ADAMTS-7, -12/COMP.

Nous proposons que de telles convergences évolutives soient favorisées par une pression de sélection visant à augmenter la complexité de la matrice extracellulaire par augmentation des interactions possibles entre les protéines matricielles. Cromar et al. (2014) [Cro+14] a démontré que de nouveaux arrangements de domaines contribuent à l'augmentation de la complexité de la matrice extracellulaire chez les vertébrés, et que malgré un nombre limité de domaines existants, explorer la combinatoire des arrangements de domaines permet l'émergence de nouvelles fonctions. D'une manière similaire, si le nombre de protéines de la matrice extracellulaire est limité, augmenter les interactions possibles des ADAMTS-TSL peut également contribuer à augmenter la complexité de la matrice. C'est ainsi que dans un contexte dans lequel la combinatoire en domaines, leur apparition et leur brassage expliquent la diversité fonctionnelle des protéines matricielles [Hyn12], notre hypothèse propose que la convergence évolutive d'interactions entre protéines matricielles, comme observé ici pour les ADAMTS, peut également constituer un mécanisme de diversification des fonctions des protéines matricielles.

8.3.3 Expansion paralogue des hyalectanases et spécificités fonctionnelles d'ADAMTS-5

La sous-famille des hyalectanases comprend sept gènes : ADAMTS-1, ADAMTS-4, ADAMTS-5, ADAMTS-8, ADAMTS-9, ADAMTS-15 et ADAMTS-20, qui dégradent des protéoglycans tels que l'aggrecan (ACAN), le versican (VCAN), le brevican (BCAN) et le neurocan (NCAN) [Fon+21]. Nous avons identifié le gène G96 comme l'ancêtre commun des gènes hylactanases auquel seraient gagnées les interactions avec ACAN et VCAN (Figure 8.11A). Les hyalectanases partagent également d'autres partenaires d'interactions, comme LRP1 qui est identifié comme partenaire de ADAMTS-4, ADAMTS-5, ADAMTS-9 et ADAMTS-20 (Figure 8.11B).

8.3.3.1 ADAMTS-5 : une hyalectanase particulière

Parmi les hyalectanases, ADAMTS-4 (aggrecanase-1) et ADAMTS-5 (aggrecanase-2) sont les acteurs majeurs de la dégradation de l'aggrecan (ACAN), le composant majeur du cartilage articulaire. En comparaison, ADAMTS-1 [San+01] et ADAMTS-9 [Som+03] sont également capables d'interagir avec l'aggrecan, mais avec une activité de dégradation bien plus faible. Pour ce qui est de l'activité de ADAMTS-4 et ADAMTS-5, l'activité de ADAMTS-5 a d'abord été montrée comme inférieure à celle de ADAMTS-4 *in vitro* [Tor+02], avant que ADAMTS-5 soit décrit comme l'acteur principal de l'activité aggrecanase en conditions physiologiques [Sta+05 ; Gla+05 ; Gen+07]. De plus, le site d'interaction avec l'aggrecan a été caractérisé comme impliquant le domaine spacer pour ADAMTS-4 [Kas+04] et comme impliquant le domaine riche en cystéines et le domaine spacer pour ce qui est d'ADAMTS-5 [Gen+07]. D'autres substrats de ADAMTS-4 et ADAMTS-5 ont été identifiés, dont certains qu'ils partagent (e.g., VCAN [San+01 ; Fou+14]) et d'autres qu'ils ne partagent pas (e.g., COMP qui est clivé par ADAMTS-4 et pas par ADAMTS-5 [Dic+03]). Ils partagent également des interactions avec d'autres protéines, comme avec TIMP3 avec laquelle ils interagissent par le biais de leurs domaines en extrémité C-terminale [ACM09], ou LRP1 par le biais de domaines différents : le domaine cystéine riche et le spacer pour ADAMTS-4, et le motif TSP1 ainsi que son spacer pour ADAMTS-5 [Yam+14]. Toutes ces observations expérimentales suggèrent une spécificité fonctionnelle de ADAMTS-5 comparé aux autres hyalectanases.

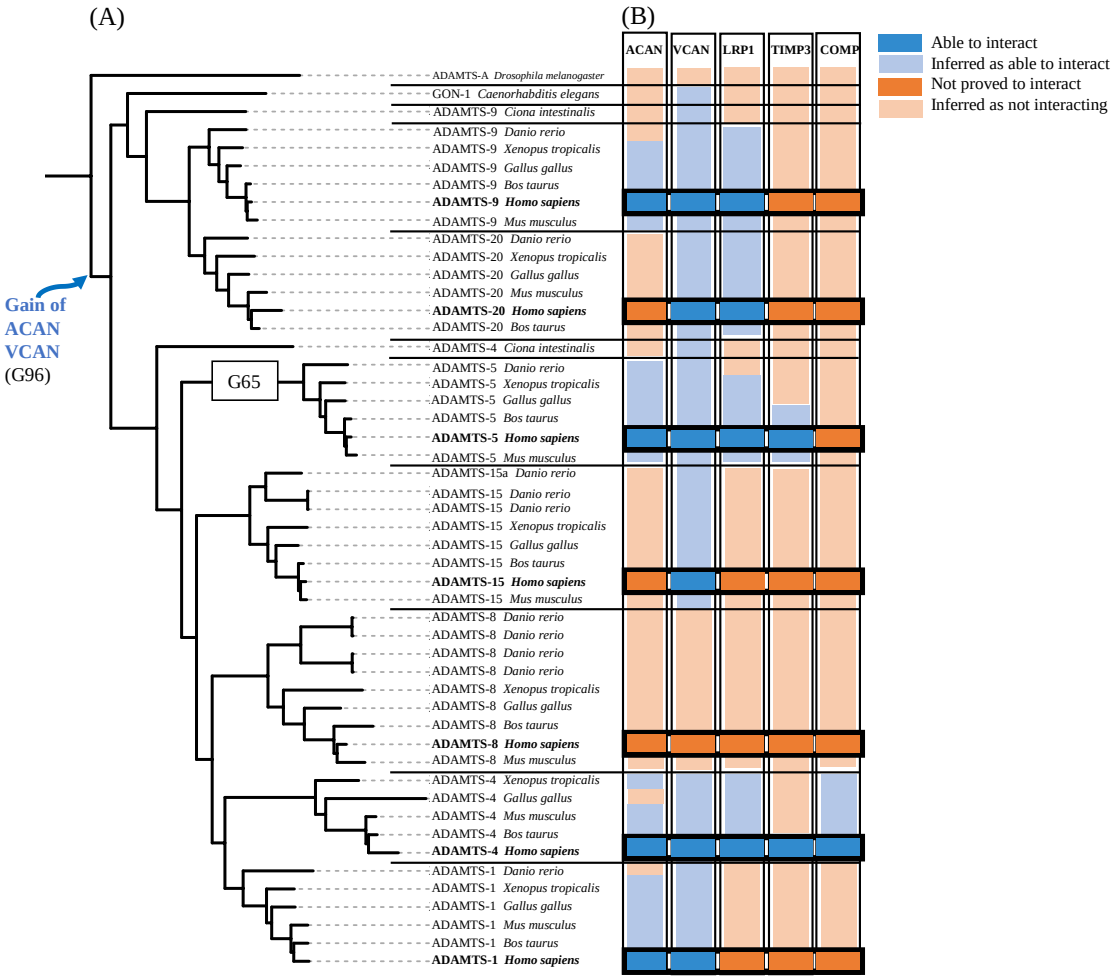

FIGURE 8.11 – **Histoires évolutives des interactions hyalectanases.**
A) Sous-arbre des hyalectanases. Le gène G96 est l'ancêtre commun des hyalectanases, et le gain des interactions ACAN et VCAN y a été inféré. B) Heatmap : les interactions publiées pour les hyalectanases humaines avec ACAN, VCAN, LRP1, TIMP3 et COMP sont représentées en bleu foncé et l'absence d'interaction est représentée en orange foncé. Les interactions inférées comme possibles par l'outil `PastML` sont représentées en bleu clair, les interactions inférées comme impossibles sont représentées en orange clair.

8.3.3.2 Caractérisation de modules fonctionnels des hyalectanases et de ADAMTS-5

C'est pourquoi, en plus des modules gagnés à l'ancêtre des hyalectanases (G96), nous nous sommes intéressés aux modules gagnés à l'ancêtre des orthologues de ADAMTS-5 qui pourraient en expliquer le caractère spécifique. Nous avons identifié G65 comme l'ancêtre commun des orthologues d'ADAMTS-5 (Figure 8.11A), auquel notre modèle identifie 12 modules qui y sont spécifiquement gagnés, dont 11 sont encore présents chez le descendant humain (boites vertes sur la Figure 8.12A). Les modules gagnés par l'ancêtre des hyalectanases (G96) sont localisés dans le domaine cystéine riche et l'extrémité N-term du domaine spacer (boites violettes sur la Figure 8.12A). Les 12 modules de G65 sont eux principalement localisés dans le propeptide, le domaine spacer et le motif TSP1 en C-term de la protéine. Ces modules sont distants dans la séquence de ADAMTS-5, mais localisent dans la même région de la structure 3D prédite par `AlphaFold` (en vert sur la Figure 8.12D), ce qui suggère l'implication de ces modules au sein d'une même interface de la protéine. Nous proposons alors ces modules gagnés à G65 comme des candidats de motifs associés aux spécificités fonctionnelles de ADAMTS-5.

La majorité des gains en modules à G96 et G65 se localisent au sein du domaine riche en cystéine et du domaine spacer, c'est pourquoi nous avons analysé plus en détails les segments des modules présents au sein de ces domaines qui sont connus pour leur implication dans l'interaction avec l'aggrecan. Nous avons alors comparé les différences de modules au sein de ces domaines cystéine riche et spacer pour toutes les hyalectanases humaines, ainsi que trois ADAMTS-TSL humaines n'appartenant pas à la sous-famille : ADAMTS-6, ADAMTS-10 (deux paralogues proches évolutivement) et ADAMTSL-4 (un paralogue distant évolutivement). Cette analyse est présentée en Figure 8.12B. Dans un premier temps, nous observons qu'une majorité des modules est partagée par toutes ces ADAMTS-TSL, indiquant une origine plus ancienne que l'ancêtre des hyalectanases G96. Les modules B773 et B781 sont partagés par les hyalectanases spécifiquement (modules gagnés à G96, leur ancêtre commun). D'une manière importante, les deux aggrecanases ADAMTS-4 et ADAMTS-5 présentent chacune des modules qui diffèrent. La protéine ADAMTS-5 est caractérisée par quatre modules B1602, B1603, B1605 et B1607 conservés spécifiquement au sein de ses orthologues (modules gagnés à G65, l'ancêtre commun des orthologues ADAMTS-5). Les deux modules B1603 et B1607 correspondent à deux des trois boucles hypervariables décrites comme médiatrices de la spécificité de l'interaction des ADAMTS [San+19b] (β3-β4 et β9-β10, respectivement). Cependant, aucun module

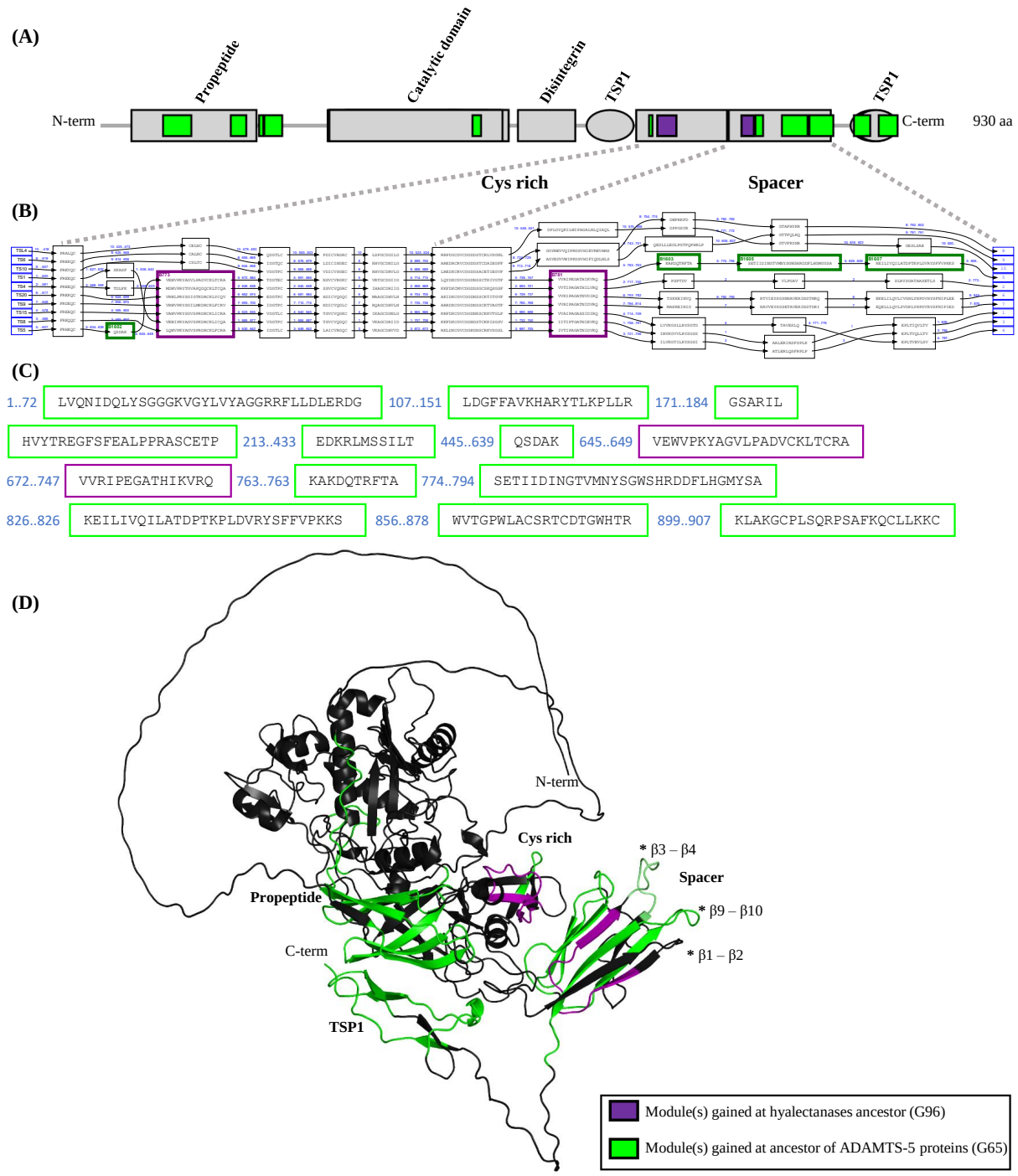

FIGURE 8.12 – **Modules gagnés à l'ancêtre des hyalectanases et de ADAMTS-5.**
A) Les segments des modules gagnés aux gènes ancestraux G65 (ancêtre des orthologues ADAMTS-5) et G96 (ancêtre des hyalectanases) sont représentés sur la séquence humaine de ADAMTS-5 (RefSeq NP_008969.2). Les domaines sont représentés en gris, les segments des modules gagnés à G65 en vert, les segments des modules gagnés à G96 en violet. **B)** Extrait du PLMA des hyalectanases humaines et de trois autres ADAMTS-TSL humaines (ADAMTS-6, ADAMTS-10 et ADAMTSL-4) des domaines cystéine riche et spacer uniquement. Chaque séquence d'ADAMTS-TSL alignée est numérotée (de 1 à 10), et les *indel* sont indiqués par des flèches noires dont le numéro de la séquence ainsi que l'intervalle entre les modules sont indiqués. Les modules gagnés à G96 sont représentés par des boites violettes, les modules gagnés à G65 sont représentés par des boites vertes, et les autres modules présents chez ces ADAMTS-TSL le sont par des boites noires. **C)** Séquences complètes des segments provenant des modules gagnés à G65 et G96 de la protéine ADAMTS-5 humaine (RefSeq NP_008969.2). **D)** Projections des modules gagnés à G65 et G96 sur la structure de ADAMTS-5 humaine prédite par `AlphaFold` (AF-Q9UNA0-F1). Les trois boucles hypervariables β1-β2, β3-β4, β9-β10 décrites par Santamaria et al., 2019 [San+19b] sont indiquées par *****.

ne correspond à la troisième boucle hypervariable β1-β2, ce qui suggère que sa séquence n'est pas conservée dans les séquences homologues.

Toutes ces données nous permettent de proposer un scénario de l'évolution du domaine spacer (Figure 8.13), comme un acteur important de la spécialisation fonctionnelle des hyalectanases/ADAMTS-5 : un premier module conservé serait acquis chez l'ancêtre des hyalectanases (748-762 dans NP_008969.2, boite violette sur la Figure 8.12B), avant l'acquisition de trois modules conservés spécifiquement chez ADAMTS-5 (764-773, 795-825 et 827-855 dans NP_008969.2, boite verte sur la Figure 8.12B). De plus, nous avons également étudié les neuf autres modules gagnés à G65 (l'ancêtre des orthologues ADAMTS-5), qui sont eux localisés hors du spacer (Figure 8.12A et C) et éloignés de ce dernier dans la séquence, mais très proches sur la structure prédite de la protéine (Figure 8.12D). Ces observations suggèrent l'implication des 12 modules gagnés à G65, comme constituant une interface complexe (la séquence des segments de ADAMTS-5 humaine est représentée en vert sur la Figure 8.12C) qui peut expliquer la spécificité fonctionnelle de ADAMTS-5.

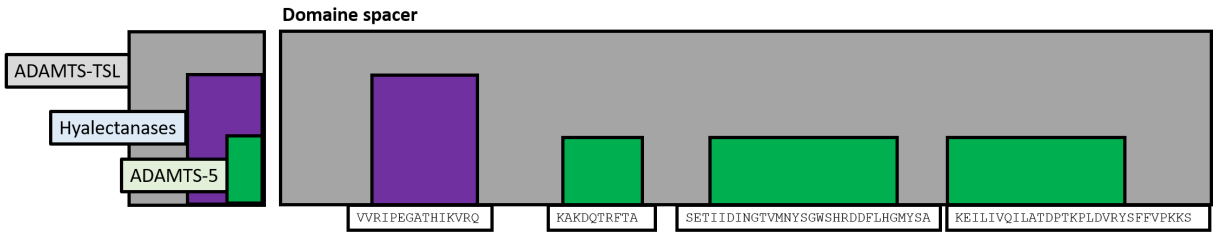

FIGURE 8.13 – **Schéma de l'évolution du domaine spacer chez les hyalectanases.**
Le domaine spacer (Pfam PF05986, *ADAM_spacer1*) commun à toutes les ADAMTS-TSL est représenté en gris, le module qui y est gagné à l'ancêtre de la sous-famille des hyalectanases (G96) est représenté en violet et les trois modules gagnés à l'ancêtre des orthologues ADAMTS-5 (G65, appartenant à la sous-famille des hyalectanases) sont représentés en vert. Les séquences des segments de ADAMTS-5 humaine pour les quatre modules sont également indiquées.

Nous prédisons donc différents modules comme étant impliqués dans des spécificités fonctionnelles au sein des hyalectanases, en particulier leurs différences d'affinité avec l'aggrecan [San+01 ; San+21 ; Som+03 ; Som+03 ; Sta+05 ; Gla+05 ; Gen+07]. L'aggrecanase physiologique ADAMTS-5 présente 12 modules spécifiques, qui bien que distants sur la séquence, sont très proches dans l'espace sur la structure prédite de la protéine, décrivant ainsi une interface spécifique d'interaction avec l'aggrecan. La spécificité fonctionnelle de l'interface que nous proposons est soutenue par le fait qu'elle contient les boucles hypervariables de ADAMTS-5 (Figure 8.12D), connues pour leurs importances dans l'activité aggrecanase [San+19a].

Depuis plus de dix ans, des anticorps monoclonaux et des petites molécules sont développés afin d'inhiber les effets délétères de ADAMTS-5 dans l'ostéoarthrite [Jia+21], cependant les effets secondaires et leur manque de spécificité empêchent toute progression vers des tests cliniques. C'est pourquoi, les modules conservés que nous proposons comme spécifiques de ADAMTS-5 pourraient être une très bonne piste pour la conception de nouvelles molécules inhibitrices.

Conclusion

Nous proposons dans cette thèse un modèle regroupant conjointement l'évolution des modules et des phénotypes de manière à étendre la méthode phylogénomique à la caractérisation de modules fonctionnels. Pour ceci, nous utilisons l'arbre des gènes de référence des ADAMTS-TSL comme modèle. De plus, nous mettons à disposition une implémentation du *workflow* complet, afin d'en permettre une application à d'autres jeux de séquences homologues, et d'autres types de phénotypes. Pour ce qui est de son application aux ADAMTS-TSL, nous mettons également à disposition une visualisation interactive sur le site `Itol`, permettant d'explorer la totalité de nos données et nos résultats. Finalement, nous avons étudié les événements de coapparition module(s)-interaction(s), afin d'associer modules(s) et interaction(s) sur la base de leur acquisition commune au cours de l'évolution des gènes, et ainsi de proposer des motifs fonctionnels impliqués dans ces interactions. Nous avons identifié 45 de ces événements, qui, confronté à la littérature, nous permettent de valider notre approche et de prédire des motifs fonctionnels.

⇒ **Implémentation du workflow PhyloCharMod :**
`https://github.com/OcMalde/PhyloCharMod_publ`

⇒ **Visualisation interactive sur le site Itol :**
`https://itol.embl.de/tree/13125419158431781652196295`

Troisième partie

Discussions et Perspectives

Chapitre 9

Discussions et Perspectives

Les méthodes classiques permettant de caractériser des protéines, leurs régions importantes et leurs fonctions sont essentiellement basées sur l'hypothèse qu'une protéine ne possède qu'une unique fonction et que toutes ses régions importantes y sont impliquées. Cependant, ces méthodes arrivent à leurs limites dans le cas de protéines complexes telles que les ADAMTS-TSL, une famille multigène, multidomaine et multifonctionnelle, présentant des points communs ainsi que des spécificités en termes de similarité de séquences et de fonctions [Moh+21a]. Ainsi, dans le cas des protéines multidomaines et multifonctionnelles, associer une région spécifique de la protéine à une fonction reste un défi. Dans ce contexte, ce projet de thèse vise à développer une nouvelle approche permettant d'identifier des régions fonctionnelles.

Nous proposons d'associer des segments de séquences conservées (non contiguës) à des fonctions, sur la base de leur évolution partagée. Il s'agit d'une extension de l'approche phylogénomique à la prédiction de régions fonctionnelles. Pour cela, nous intégrons les histoires évolutives des espèces, des gènes, des modules et des phénotypes. La modularité des protéines multidomaines est principalement due au brassage exonique [CFS10 ; Pat96]. Le phénomène évolutif à l'origine de cette complexité [Kaw+09 ; CFS10], bien qu'accepté et étudié [SOG19], reste compliqué à décrire et à modéliser par une liste d'évènements [LB19b ; Oak17 ; WRK12]. Nous avons fait le choix de modéliser l'évolution indépendante des espèces, des gènes et des modules (segments de séquences conservés) en utilisant leur arbre phylogénétique respectif qui seront ensuite interprétés grâce à une réconciliation multiModules-Gènes-Espèces basée sur un modèle DTL (duplication, transfert, perte), ceci afin de considérer l'interdépendance de ces trois niveaux d'évolution. L'analyse des résultats et des associations proposées par la réconciliation nous permet ensuite de constituer une carte des présences/absences des modules au cours de l'évolution des gènes qui va caractériser les motifs fonctionnels spécifiques à tous les sous-groupes d'ADAMTS-TSL possédant une origine évolutive commune.

Nous proposons dans cette thèse une approche cumulant approche phylogénomique et caractérisation de régions conservées à l'échelle de sous-groupes de protéines. Les approches de caractérisation de régions conservées à l'échelle de sous-famille/sous-groupe de séquences homologues classiques nécessitent une description en amont des sous-familles [DC12] ou une prédiction en un nombre restreint de sous-familles [Van+16]. En effet, l'approche standard nécessite d'identifier les séquences qui appartiennent à une même sous-famille/sous-groupe fonctionnel et qui sont alors susceptibles de présenter une conservation spécifique. Au lieu de définir une conservation à l'échelle du sous-groupe sur la comparaison des séquences de ce dernier, nous proposons de nous affranchir d'un alignement multiple des séquences et de considérer toutes les conservations locales partielles possibles (ici les modules), que ces dernières comprennent des séquences lointaines évolutivement ou non. Ceci permet d'associer des zones conservées à des sous-groupes consistants évolutivement (i.e., les bipartitions de l'arbre des gènes) tout en considérant la modularité des protéines (i.e., une région n'est pas nécessairement héritée depuis un ancêtre direct). Certains de ces clades sont également associés à des fonctions.

Appliquer cette approche à 214 séquences de protéines ADAMTS-TSL de 9 espèces (incluant un groupe externe de protéines ADAM) nous a permis de proposer une carte des acquisitions et des pertes de modules et d'interactions protéine-protéine au cours de l'évolution des ADAMTS-TSL de la lignée humaine. Cette carte nous a permis de caractériser 186 signatures de modules spécifiques à des sous-groupes de séquences, dont 45 sont associées à des interactions protéine-protéine. L'étude et la validation de ces associations « gain de modules - gain de PPI », nous a permis de mettre en lumière des processus évolutifs qui expliquent l'hétérogénéité fonctionnelle de certaines protéines ADAMTS-TSL humaines.

Cromar et al. (2014) [Cro+14] a démontré que le réarrangement d'un nombre limité de domaines contribue à l'augmentation de la complexité de la matrice extracellulaire chez les vertébrés, de nouvelles combinaisons de domaines permettant l'émergence de nouvelles fonctions. Nous proposons ici l'hypothèse de la convergence évolutive d'interactions comme mécanisme de l'évolution de la matrice extracellulaire. Ainsi, un même partenaire d'interaction impliquera différentes régions chez différentes protéines homologues (e.g., les PPI ADAMTS-COMP/CCN2). Ce phénomène augmente la complexité de l'interactome de la matrice extracellulaire.

Nous mettons également en lumière la variabilité du propeptide tel que mise en évidence par Demidyuk et al. (2010) [Dem+10], ainsi que plus généralement la variabilité des extrémités C-term/N-term des protéines ADAMTS-TSL. Nous y identifions la majorité des modules conservés, parfois partagés depuis des ancêtres de paralogues. Ces acquisitions de modules conservés dans les extrémités C-term/N-term, de manière synchrone avec l'acquisition de PPI, nous permet de proposer l'hypothèse selon laquelle la variabilité des séquences aux extrémités C-term/N-term jouerait un rôle important dans l'hétérogénéité fonctionnelle des protéines ADAMTS-TSL humaines.

Finalement, nous proposons des signatures composées de modules conservés, non contiguës le long des protéines, composées de modules plus courts et plus conservés que les domaines et plus à même d'expliquer l'hétérogénéité fonctionnelle de différentes protéines et des différents sous-groupes. Nous en avons démontré leur pertinence pour le cas ADAMTS-5, et tout particulièrement par la caractérisation d'une interface d'interaction composée de modules non contigus sur la séquence, mais proches dans l'espace, selon la prédiction d'AlphaFold. La signature identifiée est spécifique à l'ancêtre des orthologues d'ADAMTS-5 et indique notamment une conservation forte de régions de son spacer.

La discrimination des différentes ADAMTS-TSL (ici illustrée pour ADAMTS-5) par des signatures s'inscrit dans une logique d'identification de nouvelles cibles thérapeutiques. En effet, les anticorps monoclonaux et les petites molécules généralement développés dans le but d'inhiber les effets délétères de ADAMTS-5 dans l'ostéoarthrite manquent de spécificité, présentant alors un grand nombre d'effets secondaires [Jia+21]. Ceci empêche toute progression de ces anticorps et molécules vers des tests cliniques. Dans le but de développer des approches plus spécifiques basées sur des cibles thérapeutiques plus précises, nous proposons un nombre limité de modules candidats, où chacun des modules peut être une cible à tester. De plus, la considération d'espèces non humaines dans nos prédictions permet de déterminer d'autres organismes chez lesquels les modules candidats seraient présents, et donc chez lesquels un ciblage peut être testé.

Un des points clés d'amélioration des prédictions fonctionnelles passe par l'acquisition de nouvelles connaissances. En effet, la découverte de nouvelles PPI et l'utilisation de différentes annotations (e.g., mutations, implication dans une pathologie) permettrait d'aller encore plus loin dans la caractérisation des protéines ADAMTS-TSL, et de pouvoir associer et préciser des phénotypes dans lesquels nos signatures en modules gagnés seraient impliquées. Notamment, une information intéressante, mais quasiment jamais décrite, serait la description de la non-capacité d'une protéine à effectuer une PPI, de manière

à augmenter très fortement la puissance prédictive et la précision de notre approche. Il s'agit d'interpréter les absences d'interaction (PPI). La connaissance plus approfondie de l'état (présence/absence) des PPI aux feuilles de l'arbre donnerait la possibilité à `PastML` de reconstruire des scénarios ancestraux plus complets et robustes.

Notre approche et les résultats qu'elle propose sont donc amenés à évoluer avec l'acquisition de nouvelles connaissances expérimentales, permettant de compléter les informations phénotypiques utilisées. En attendant, les signatures de modules gagnés, spécifiques à chaque sous-groupe de séquences homologues de notre carte, proposent de nombreuses signatures à explorer dans le cadre de problématiques ciblées, telles que : « qu'est-ce qui différencie ADAMTS-2, -3, -14 des autres ? », « quelles régions sont spécifiques à ADAMTS-13 ? », « à ADAMTSL-1, -3 ? », « à ADAMTSL-2, -4 ? », etc. Toutes ces informations sont disponibles et explorables par la communauté grâce aux vues interactive de notre **Arbre Itol**.

Prise en compte des isoformes pour améliorer les prédictions

Nous avons développé une approche de modélisation de la totalité des exons des isoformes d'un gène (i.e., graphe de segments) qu'il serait intéressant de considérer dans la totalité de notre approche. Au lieu de considérer la séquence la plus longue d'un gène, l'utilisation du graphe de segments de ce gène permettrait de prendre en compte la totalité du contenu en exons. Ceci nécessiterait d'adapter les outils qui prennent des séquences en entrées (e.g., les outils d'alignements), ou d'utiliser des outils de construction de graphes d'épissages multi-espèces tel que `ThorAxe` [Zea+21]. Cependant, l'intérêt d'une telle perspective repose également sur la connaissance d'annotations spécifiques aux différentes isoformes, qui sont aujourd'hui rarement disponibles.

Les pertes de modules et de fonctions comme source d'information

Une autre perspective intéressante est l'étude des autres scénarios qui corrèlent évolution des modules et évolution des phénotypes. Nous nous sommes intéressés aux évènements de coapparitions, mais notre modèle permet également d'identifier des évènements de co-pertes (i.e., perte synchrone de modules et d'interactions), ou des évènements plus complexes comme les modules acquis quand une interaction est perdue, ou les modules perdus quand une interaction est acquise. En effet, la perte de modules corrélée à des chan-

gements fonctionnels de la protéine, bien que plus complexe à interpréter, est également une information qu'il serait intéressante d'étudier dans le but de comprendre l'association séquence/fonction. Il s'agit là de pouvoir interpréter les pertes de modules et d'interactions au cours de l'évolution.

Généralisation de la méthode

La construction de notre jeu d'interactions protéine-protéine, à la fois automatique (extraction via l'interface `PSICQUIC`) et manuelle (étude bibliographique) est une contribution importante de notre étude des protéines matricielles ADAMTS-TSL, mais c'est également le goulet d'étranglement pour appliquer notre approche à d'autres familles. L'approche que nous avons développée au cours de cette thèse est parfaitement utilisable sur d'autres jeux de séquences homologues multidomaines. La détection des modules et la reconstruction des compositions ancestrales en modules, qui ne nécessite aucun *a priori*, est tout particulièrement généralisable. Cependant, il est nécessaire d'avoir une description de phénotypes associés à chacune de ces séquences afin d'associer modules et phénotypes. Une solution pour étudier les interactions protéine-protéine d'un autre jeu de séquences homologues serait de se limiter à une extraction automatique via `PSICQUIC`, qui pourrait également être cumulée à une extraction automatique des interactions de la base de données `STRING` [Szk+19] afin de constituer un jeu d'interactions au sens plus large (e.g., coexpression, cooccurrence dans le texte). L'utilisation d'un jeu d'interactions protéine-protéine pose le problème de l'espace de recherche mais aussi du biais de connaissances. En effet, certaines protéines sont bien plus étudiées que d'autres et les méthodes de *screening* apportent une grande quantité d'informations pour certaines protéines, déséquilibrant l'annotation des groupes comme le montre le graphe des PPI (Figure 7.4). De même, les ADAMTS-TSL de la plupart des espèces sont très peu caractérisées comparé au ADAMTS-TSL humaines, même pour une espèce modèle telle que la souris. Nous avons en conséquence fait le choix de n'utiliser que les annotations fonctionnelles des ADAMTS-TSL humaines. Cependant, disposer d'informations pour les protéines orthologues chez les autres espèces améliorerait énormément les prédictions de notre méthode.

Notre approche est en théorie applicable à n'importe quel trait sous pression de sélection, il est donc envisageable de décrire comme phénotypes divers caractères. Une possibilité serait d'utiliser les termes GO comme annotations fonctionnelles, où chaque terme GO correspondrait à un trait phénotypique, ce qui pourrait permettre de décrire automatiquement les fonctions d'un jeu de séquences homologues d'une manière homo-

gène, bien que plus abstraite. Dans la même lignée, il serait très intéressant de représenter les interactions protéine-protéine d'une manière plus abstraite, en regroupant les partenaires d'interactions similaires sous un même trait phénotypique. Par exemple, pour les ADAMTS-TSL, il serait intéressant et envisageable de regrouper les interactions avec les hyalectans ACAN, VCAN et BCAN sous un même trait phénotypique « interaction avec un hyalectan ». Cette abstraction des fonctions pourrait alors permettre de décrire des modules qui leur sont associés, moins précise que l'interaction avec un unique partenaire, mais générale à une catégorie d'interactions. Le regroupement et l'abstraction des partenaires similaires pourrait alors s'automatiser sur la base de leur homologie par le biais de leurs orthogroupes : les protéines appartenant à un même orthogroupe représenteraient alors un unique trait phénotypique. Nous pourrions envisager une généralisation de notre approche à la totalité des orthogroupes que nous avons calculés, et dont chaque trait phénotypique correspond à la capacité des différentes protéines homologues à interagir avec une protéine appartenant à un autre orthogroupe.

Propagation d'annotations en dehors de l'arbre

En plus de l'annotation par propagation réalisée dans l'arbre des gènes, nous avons également exploré la capacité de nos signatures de modules à prédire de nouvelles séquences possédant la même fonction, via le scan de séquences non présentes dans l'arbre, à la recherche de nos signatures de modules prédites. Pour ceci, nous avons représenté les signatures de modules sous forme d'un automate, nommé *protomaton* que nous pouvons aligner avec une séquence cible à l'aide de l'outil `protomatch` de la suite `Protomata` [Ker08]. Nous avons cherché la présence de nos signatures de modules au sein de différents jeux de protéines : protéome humain, Uniprot nr90 et différents jeux composés uniquement des séquences des autres protéines qui sont également capables d'interagir avec les partenaires d'interactions associés à nos signatures de modules. Nos résultats préliminaires révèlent une très forte spécificité de nos signatures de modules pour les protéines dont elles proviennent, mais également pour leurs orthologues non compris dans notre modèle, ce qui suggère une très forte spécificité des signatures de modules pour les sous-groupes qui descendent de l'ancêtre qu'elles décrivent. Ces résultats préliminaires ouvrent ainsi la perspective de l'utilisation de signatures de modules, au format de protomates, comme outil de propagation d'annotations fonctionnelles à des séquences de protéines non annotées en dehors de l'arbre.

Conclusion finale

Pour conclure, nous proposons une approche innovante de recherche de motifs fonctionnels au sein de familles de protéines homologues, et tout particulièrement dans le cadre de protéines multidomaines. Cette approche permet l'identification de motifs conservés, fins, non contigus, à même de discriminer les spécificités des sous-groupes de protéines homologues. Mais nous allons plus loin que l'identification de motifs discriminants, nous proposons également une approche phylogénomique dans le but d'associer des fonctions à ces motifs. Pour ceci, nous proposons de reconstruire conjointement l'histoire évolutive des espèces, des gènes, des modules ainsi que des phénotypes afin de décrire des associations génotypes/phénotypes sur la base de leur évolution concomitante.

⇒ **Implémentation du workflow PhyloCharMod :**
`https://github.com/OcMalde/PhyloCharMod_publ`

⇒ **Visualisation interactive sur le site Itol :**
`https://itol.embl.de/tree/13125419158431781652196295`

Chapitre 10

Annexes

10.1 Analysis of the uncertainty in the ADAMTS-TSL phylogenetic tree

Appendix of : *Phylogenetic prediction of functional modules in ADAMTS-TSL proteins reveals sequence signatures involved in protein-protein interactions*

Analysis of the uncertainty in the ADAMTS-TSL phylogenetic tree

Olivier Dennler[1,2], François Coste[1], Samuel Blanquart[1], Catherine Belleannée[1], Nathalie Théret[1,2*],

1 Univ Rennes, Inria, CNRS, IRISA, UMR 6074, Rennes, France
2 Univ Rennes, Inserm, EHESP, Irset, UMR S1085, Rennes, France

* nathalie.theret@univ-rennes1.fr

As ADAMTS-TSL proteins are composed of multiple domains, with many motif and domain repeats, we are cautious about the inferred phylogeny (i.e., the reference tree, Fig 3 in main).

As described in the Materials and Methods, the gene phylogeny was inferred from the 214 representative protein sequences, each being encoded by a single gene (173 ADAMTS and ADAMTSL as the in-group, 41 ADAM as the out-group). This calculation required several steps and resulted in the final phylogeny hereafter referred to as the *reference gene tree*. As a first step, we used the multiple sequence aligner PASTA [1] (default settings) to compute the multiple alignment of the 214 proteins and the maximum likelihood phylogeny inference software RAxML [2] (default settings) to infer a first initial phylogeny. This initial tree was corrected using Treefix [3] taking into account the species tree as defined in the NCBI Taxonomy, thus possibly modifying the topology of the initial tree. The branch lengths (lost by Treefix) were then recomputed using PhyML [4] and the resulting tree was rooted using ADAM proteins as the out-group, resulting in the final phylogeny, designated as the *reference gene tree* (Fig 3 in main).

To analyse the uncertainty resulting from the taxon and gene sampling, and to identify robust bi-partitions in our *reference gene tree*, we analyzed different ADAMTS-TSL protein samples. Our study included two types of re-sampling: protein sequence from the **out-group** (i.e., ADAM re-sampling) and **in-group** (i.e., ADAMTS-TSL re-sampling) of the reference gene tree.

The out-group gene re-sampling used subsets of the out-group ADAM proteins. We constructed 6 different data-sets, each including all the 125 ADAMTS and 48 ADAMTSL sequences plus a subset of the 41 ADAM out-group sequences including either i) ADAM9 (tree in Fig 1), ii) ADAM12 (tree in Fig 2), iii) ADAM17 (tree in Fig 3), iv) ADAM10 (tree in Fig 4), v) ADAM9 plus ADAM12 (tree in Fig 5) and vi) ADAM17 plus ADAM10 (tree in Fig 6). Then, we constructed the multiple sequence alignment (MSA) using PASTA software [1]and inferred the phylogeny of each data-set using the RAxML software [2]. ADAM out-groups were used to root the trees. Finally we computed the "support values" of the reference gene bipartitions, indicating, for each bipartition, the frequency for the resampled trees among 6 that have the bipartition (Fig 7).

The in-group gene re-sampling used ADAMTS-TSL proteins from different species. We constructed an alternative ADAMTS-TSL data-set, composed of 255 known sequences (canonical Uniprot) from six species not considered in our reference data-set (*Rattus norvegicus, Xenopus laevis, Canis lupus familiaris, Felis catus, Anolis carolinensis, Pan troglodytes*). Next, we constructed 20 different data-sets, each containing all 173 ADAMTS-TSL sequences from the reference data-set plus 10 sequences randomly chosen from the alternative data-set. Then, we constructed the multiple sequence alignment (MSA) using PASTA software [1] and inferred the phylogeny of each data-set using the RAxML software [2]. Finally we

computed the "support values" of the reference gene tree bipartitions, using these 20 re-sampled trees pruned from the added sequences (Fig 8).

All trees were visualized using the Interactive Tree Of Life software, Itol [5].

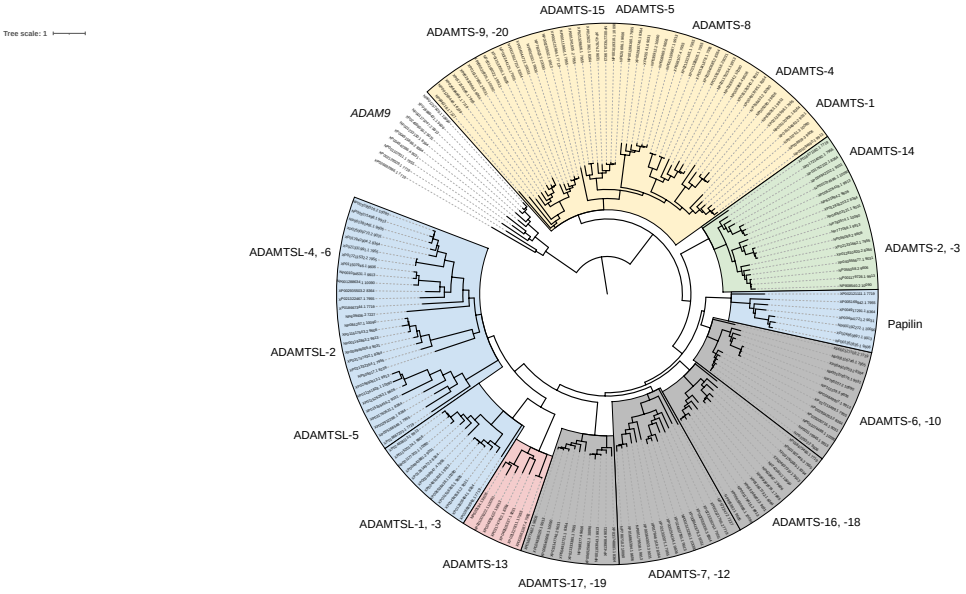

Fig 1. Gene tree with ADAM9 as the out-group The 125 ADAMTS, 48 ADAMTSL and 10 ADAM9 sequences were aligned using the PASTA software with default parameters and the phylogeny was inferred using the RAxML software with default parameters. The figure was performed using the Tree Of Life software, Itol

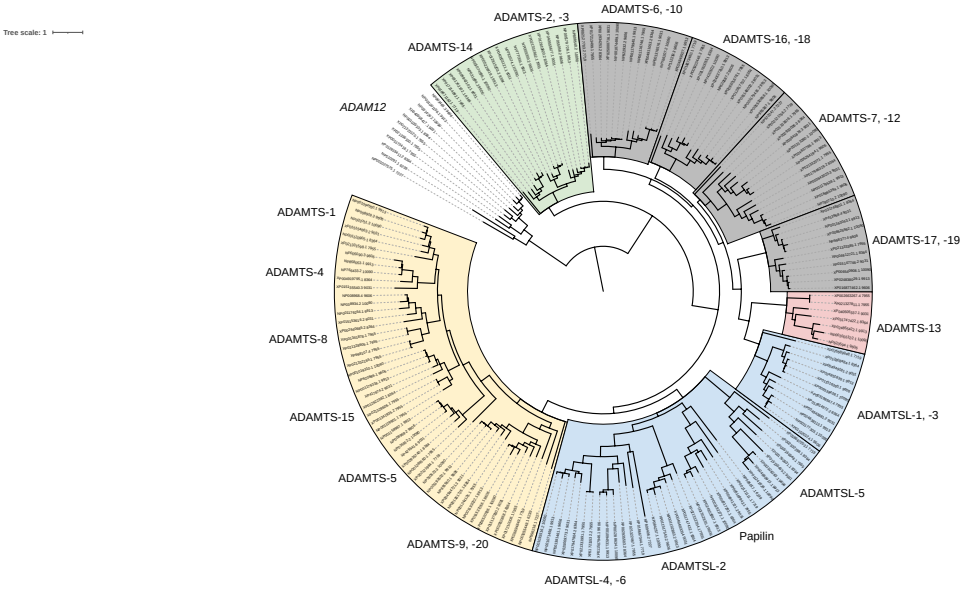

Fig 2. Gene tree with ADAM12 as the out-group The 125 ADAMTS, 48 ADAMTSL and 11 ADAM12 sequences were aligned using the PASTA software with default parameters and the phylogeny was inferred using the RAxML software with default parameters. The figure was performed using the Tree Of Life software, Itol

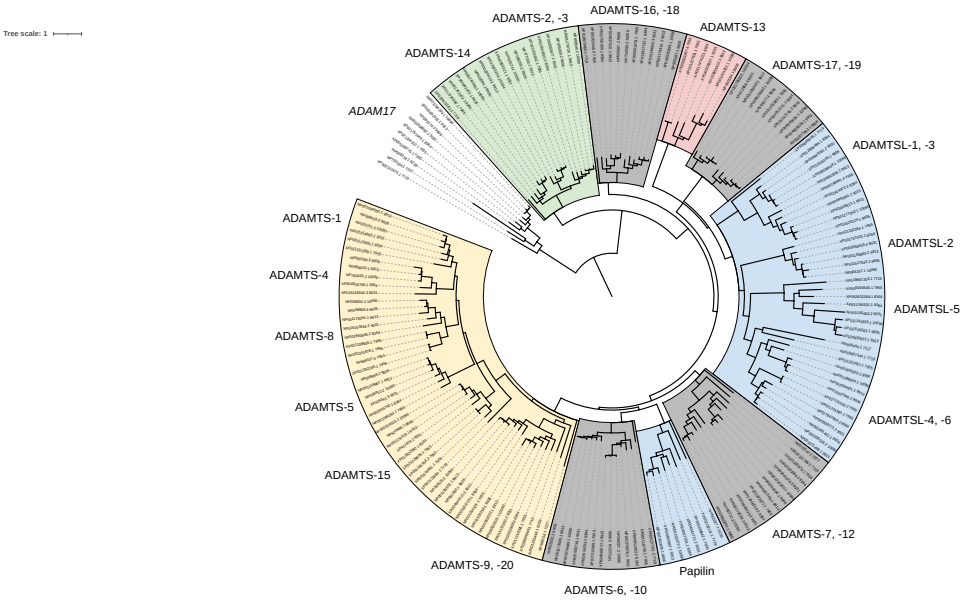

Fig 3. Gene tree with ADAM17 as the out-group The 125 ADAMTS, 48 ADAMTSL and 10 ADAM17 sequences were aligned using the PASTA software with default parameters and the phylogeny was inferred using the RAxML software with default parameters. The figure was performed using the Tree Of Life software, Itol

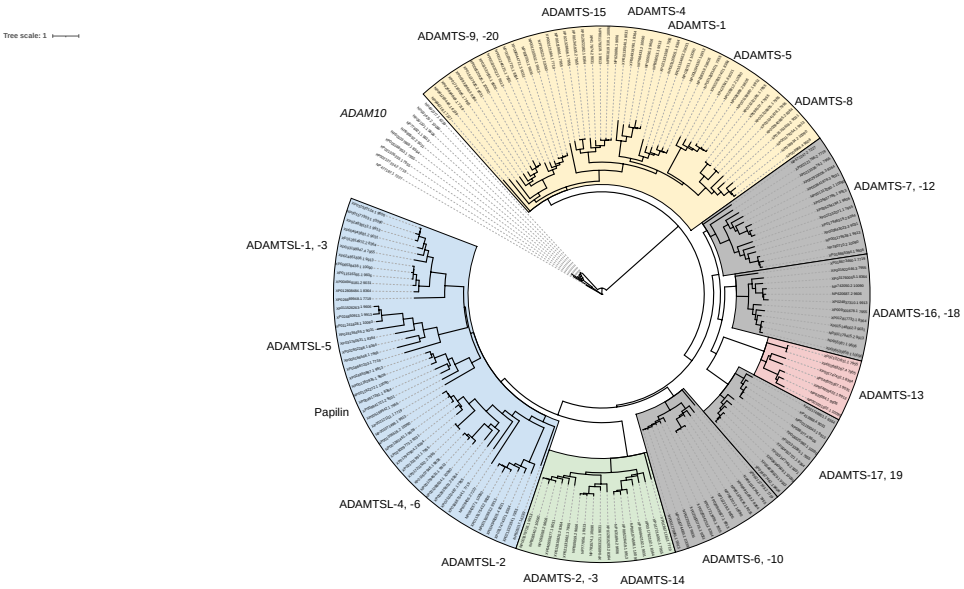

Fig 4. Gene tree with ADAM10 as the out-group The 125 ADAMTS, 48 ADAMTSL and 10 ADAM10 sequences were aligned using the PASTA software with default parameters and the phylogeny was inferred using the RAxML software with default parameters. The figure was performed using the Tree Of Life software, Itol

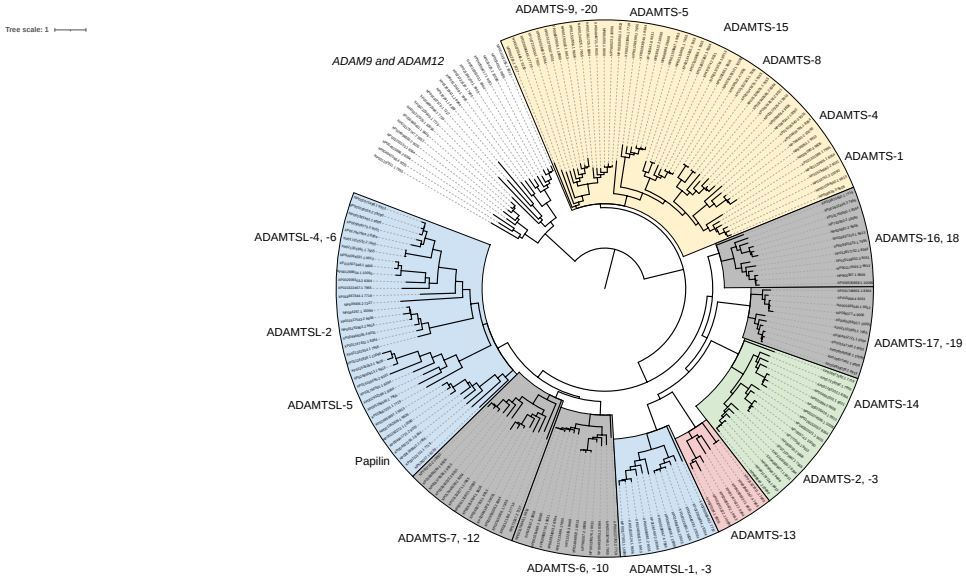

Fig 5. Gene tree with ADAM9 and ADAM12 as the out-group The 125 ADAMTS, 48 ADAMTSL, 10 ADAM9 and 11 ADAM12 sequences were aligned using the PASTA software with default parameters and the phylogeny was inferred using the RAxML software with default parameters. The figure was performed using the Tree Of Life software, Itol

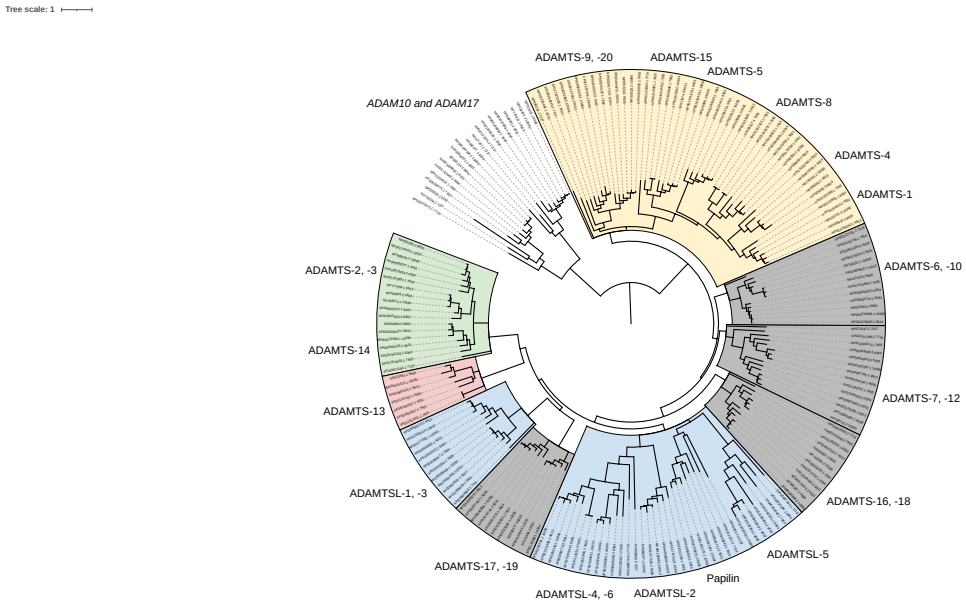

Fig 6. Gene tree with ADAM10 and ADAM17 as the out-group The 125 ADAMTS, 48 ADAMTSL, 10 ADAM10 and 11 ADAM17 sequences were aligned using the PASTA software with default parameters and the phylogeny was inferred using the RAxML software with default parameters. The figure was performed using the Tree Of Life software, Itol

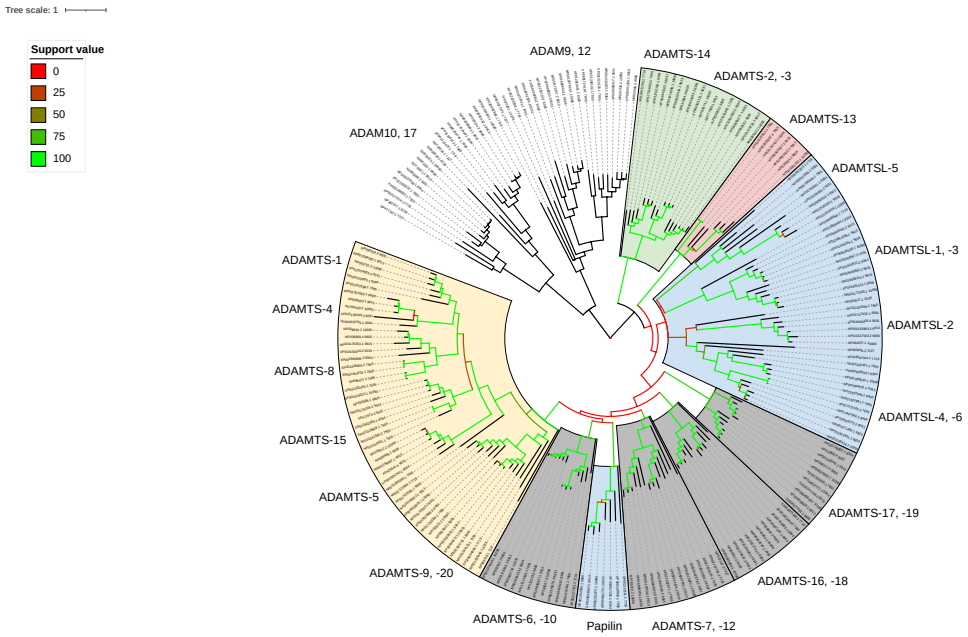

Fig 7. Reference gene tree with support values from the out-group re-sampling trees The out-group gene re-sampling used subsets of the out-group ADAM proteins. We constructed 6 different data-sets, each including all the 125 ADAMTS and 48 ADAMTSL sequences plus a subset of the 41 ADAM out-group sequences including either i) ADAM9, ii) ADAM12, iii) ADAM17, iv) ADAM10, v) ADAM9 plus ADAM12 and vi) ADAM17 plus ADAM10. Then, we constructed the multiple sequence alignment (MSA) and inferred the phylogeny of each data-set. Finally we computed support values on the reference gene tree (branches color scale, from red for 0, to green for 100), using these re-sampled trees. The figure was performed using the Tree Of Life software, Itol

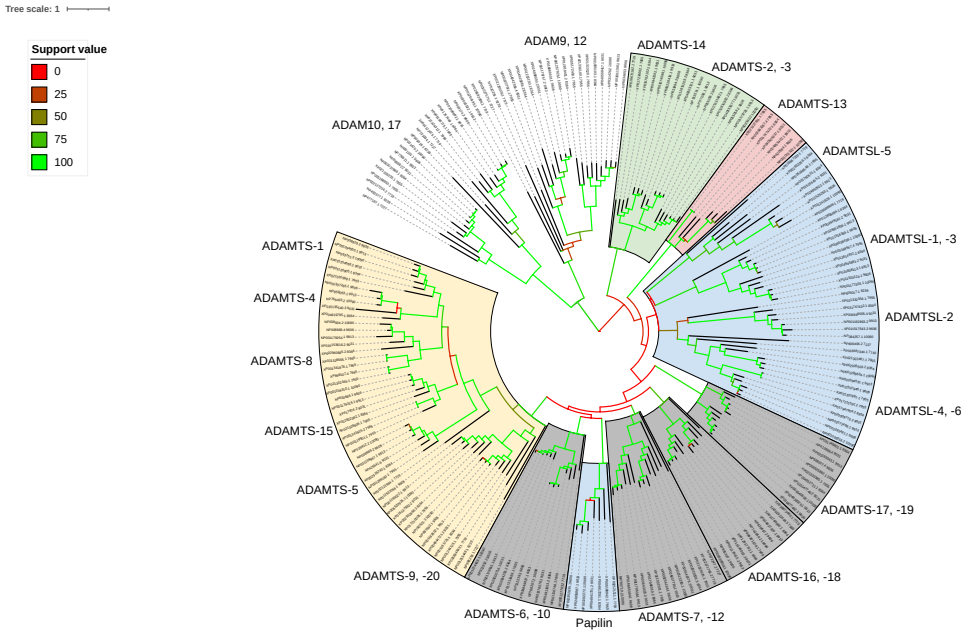

Fig 8. Reference gene tree with support values from the in-group re-sampling trees The in-group gene re-sampling used ADAMTS-TSL proteins from different species. We constructed an alternative ADAMTS-TSL data-set, composed of 255 known sequences (canonical Uniprot) from six species not considered in our reference data-set (*Rattus norvegicus, Xenopus laevis, Canis lupus familiaris, Felis catus, Anolis carolinensis, Pan troglodytes*). Next, we constructed 20 different data-sets, each containing all 173 ADAMTS-TSL sequences from the reference data-set plus 10 sequences randomly chosen from the alternative data-set. Then, we constructed the multiple sequence alignment (MSA) and inferred the phylogeny of each data-set. Finally we computed support values on the reference gene tree (branches color scale, from red for 0, to green for 100), using these re-sampled trees. The figure was performed using the Tree Of Life software, Itol

References

1. Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. Journal of Computational Biology. 2015;22(5):377–386. doi:10.1089/cmb.2014.0156.

2. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–1313. doi:10.1093/bioinformatics/btu033.

3. Wu YC, Rasmussen MD, Bansal MS, Kellis M. TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees. Systematic Biology. 2013;62(1):110–120. doi:10.1093/sysbio/sys076.

4. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010;59(3):307–321. doi:10.1093/sysbio/syq010.

5. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research. 2019;47(W1):W256–W259. doi:10.1093/nar/gkz239.

10.2 Browsing the ADAMTS-TSL Itol tree

Appendix of : *Phylogenetic prediction of functional modules in ADAMTS-TSL proteins reveals sequence signatures involved in protein-protein interactions*

Browsing the ADAMTS-TSL Itol tree

Olivier Dennler[1,2], François Coste[1], Samuel Blanquart[1], Catherine Belleannée[1], Nathalie Théret[1,2*],

1 Univ Rennes, Inria, CNRS, IRISA, UMR 6074, Rennes, France
2 Univ Rennes, Inserm, EHESP, Irset, UMR S1085, Rennes, France

* nathalie.theret@univ-rennes1.fr

1 Navigating the tree

An important contribution of this work is the availability of all the data through an interactive tree (automatically generated at the end of the pipeline), using the Interactive Tree Of Life software, Itol [1]. The ADAMTS-TSL Itol tree is available Here. When accessing the Itol tree, the user will be presented with the original view of our ADAMTS-TSL tree (Fig 1), allowing him to navigate in the tree, to modify the representation, or to activate different datasets.

An HTML popup window (Fig 2) provides detailed information about each gene node, including the name of the gene, the nature of the gene (ancestor or leaf), the PPIs associated with the gene and PPI gain/loss with respect to the ancestor, the module composition of the gene and module gain/loss with respect to the ancestor.

Saved views (Fig 3) allow quick access to visualizations made beforehand to study cases of interest (e.g., the study cases of the article).

2 Datasets usage

The Itol tree also provides annotations as datasets, including the number, the composition and the transfer of modules, the domain composition, the speciations events and the presence of PPI. Fig 4 illustrates the "Module number" and the "Func Annotations" datasets.

In particular, the "Module composition" and the "Pfam" datasets visualize module composition (Fig 5) and domain composition (Fig 6) for each leaf, with all modules/domains information available on popups.

Each gene has its own module signature annotations (1 dataset per gene node) for both present (Fig 7) and gained (Fig 8) modules. All module signatures can be visualized with their domain contexts by first enabling the domain composition annotations (Pfam dataset).

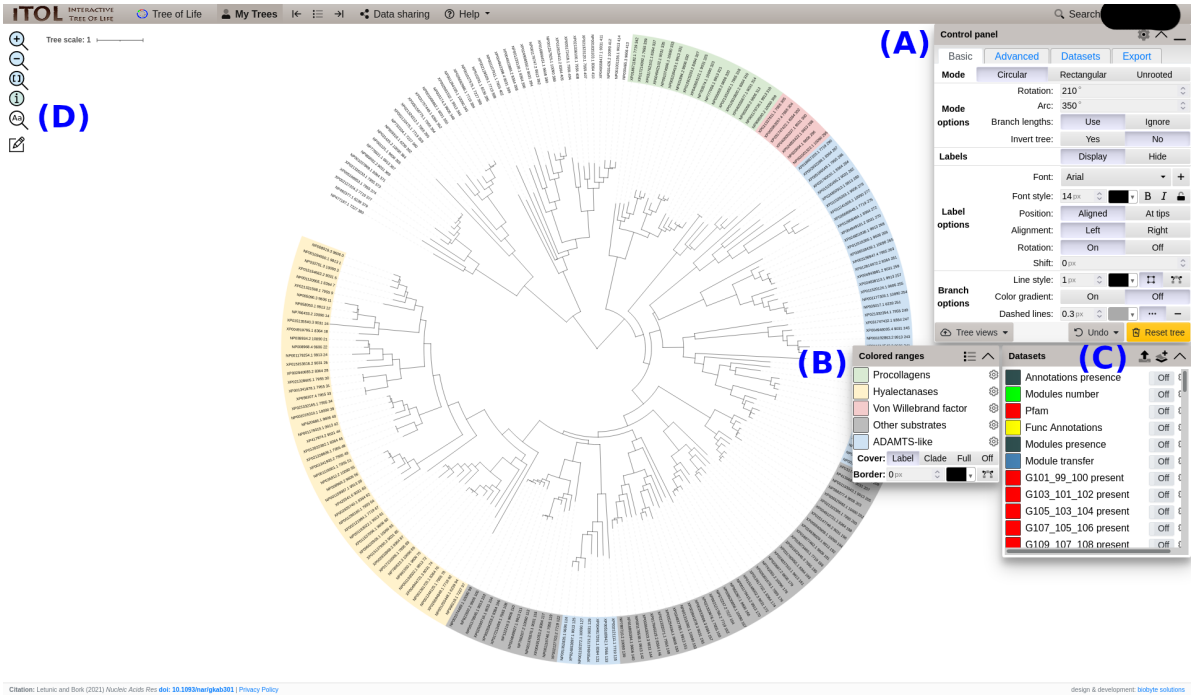

Fig 1. Original view of the ADAMTS-TSL Itol tree

(A) Itol control panel, (B) Colored ranges panel, (C) Datasets activation panel, and (D) Search tree node engine.

Fig 2. Node popup

(A) Each gene node (ancestrals and leaves) has a custom popup containing all protein, module(s) and PPI(s) information. (B) Protein information : node name (with RefSeq ID for leaves) and link to the protein entry on the NCBI website. (C) List of annotations (here PPI) present, gain and lost at this gene node. (D) List of modules present, gain and lost at this gene node.

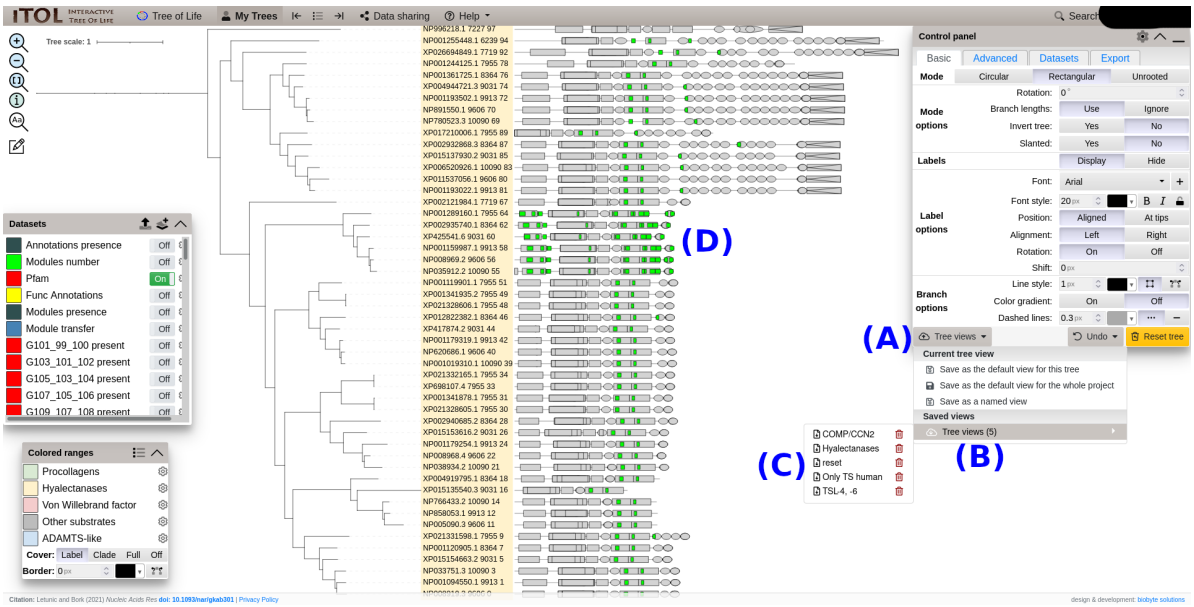

Fig 3. Saved views

(A) Tree views panel. (B) Saved views / Tree views panel. (C) List of our customs views (e.g., the hyalectanases pruned subtree and the corresponding modules signatures). (D) The hyalectanases saved view.

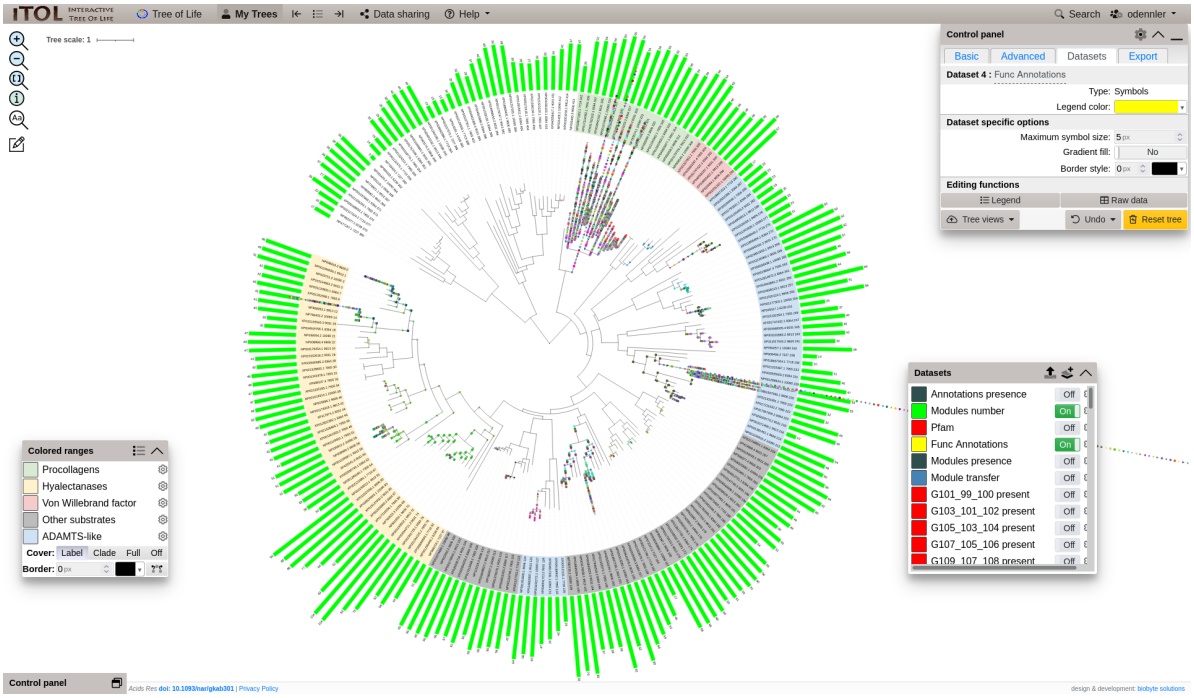

Fig 4. Example of datasets

Both the "Module number" (number of modules at each leaf) and "Func Annotations" (PPIs presences are symbolised with combinations of shapes/colors) datasets are enabled on the default view.

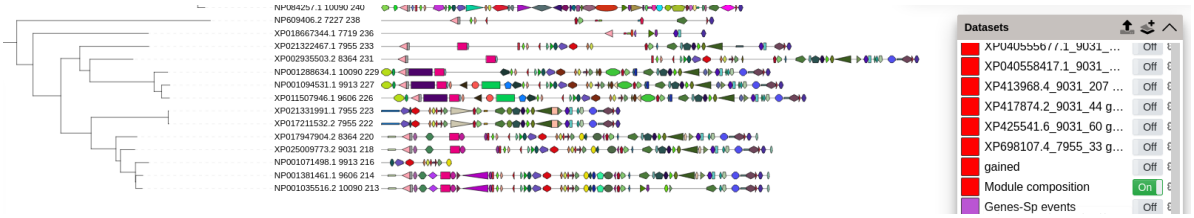

Fig 5. Module composition dataset
All leaf module compositions are represented by a mosaic of modules. Each module is a combination of shape/color and has a popup with its name and positions on the protein sequence.

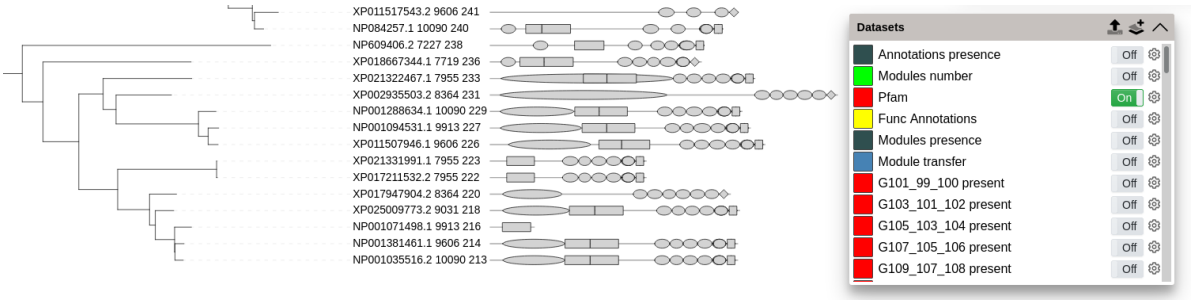

Fig 6. Pfam dataset
All leaf domain compositions are represented by grey shapes. Each domain has a popup with its ID and positions on the protein sequence.

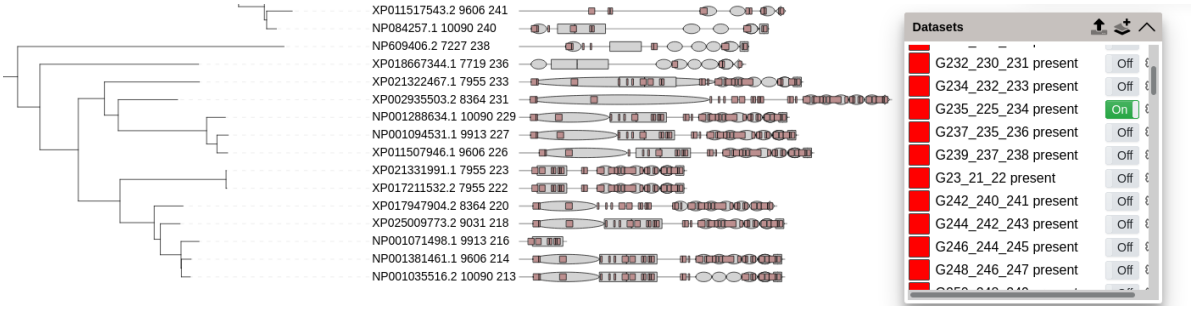

Fig 7. Example of ancestral module presence dataset
The "Pfam dataset" has been enabled prior to the "G235_225_224_present" dataset. All modules present (module composition) at the G235 ancestral gene node are represented as brown boxes on the actual proteins (leaves) where they are present.

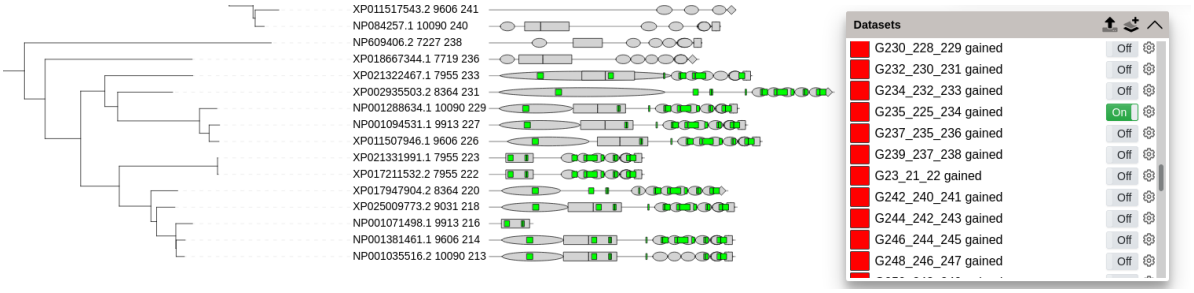

Fig 8. Example of ancestral module gained dataset
The "Pfam dataset" has been enabled prior to the "G235_225_224_gained" dataset. All modules gained (present but absent in its ancestor) at the G235 ancestral gene node are represented as green boxes on the actual proteins (leaves) where they are present.

References

1. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research. 2019;47(W1):W256–W259. doi:10.1093/nar/gkz239.

BIBLIOGRAPHIE

[Abu+17] Abulikemu ABUDUKELIMU et al., « Learning to read and write in evolution : from static pseudoenzymes and pseudosignalers to dynamic gear shifters », en, in : *Biochemical Society Transactions* 45.*3* (juin 2017), p. 635-652, ISSN : 0300-5127, DOI : 10.1042/BST20160281, URL : /biochemsoctrans/article/45/3/635/66931/Learning-to-read-and-write-in-evolution-from (visité le 28/10/2019) (cf. p. 44).

[ACM09] Alaa ABOU-ELHAMD, Oliver COOPER et Andrea MÜNSTERBERG, « Klhl31 is associated with skeletal myogenesis and its expression is regulated by myogenic signals and Myf-5 », eng, in : *Mechanisms of Development* 126.*10* (oct. 2009), p. 852-862, ISSN : 1872-6356, DOI : 10.1016/j.mod.2009.07.006 (cf. p. 224).

[Ada72] III ADAMS Edward N., « Consensus Techniques and the Comparison of Taxonomic Trees », in : *Systematic Biology* 21.*4* (déc. 1972), p. 390-397, ISSN : 1063-5157, DOI : 10.1093/sysbio/21.4.390, eprint : https://academic.oup.com/sysbio/article-pdf/21/4/390/4567482/21-4-390.pdf, URL : https://doi.org/10.1093/sysbio/21.4.390 (cf. p. 135).

[Adr+19] J. M. ADRIAN-SEGARRA et al., « Identification of Functional Protein Regions Through Chimeric Protein Construction », in : *J Vis Exp 143* (jan. 2019) (cf. p. 40).

[Aeb+18] R. AEBERSOLD et al., « How many human proteoforms are there ? », in : *Nat Chem Biol* 14.*3* (fév. 2018), p. 206-214 (cf. p. 23).

[Aga+18] R. AGARWALA et al., « Database resources of the National Center for Biotechnology Information », in : *Nucleic Acids Res* 46.*D1* (jan. 2018), p. D8-D13 (cf. p. 29).

[AKR19] R. C. ANAFI, M. S. KAYSER et D. M. RAIZEN, « Exploring phylogeny to find the function of sleep », in : *Nat Rev Neurosci* 20.*2* (fév. 2019), p. 109-116 (cf. p. 56).

[Alt+90] S. F. Altschul et al., « Basic local alignment search tool », in : *J Mol Biol* 215.*3* (oct. 1990), p. 403-410 (cf. p. 66).

[Alt+97] S. F. Altschul et al., « Gapped BLAST and PSI-BLAST : a new generation of protein database search programs », in : *Nucleic Acids Res* 25.*17* (sept. 1997), p. 3389-3402 (cf. p. 67).

[Apt20] Suneel S. Apte, « ADAMTS Proteins : Concepts, Challenges, and Prospects », eng, in : *Methods in Molecular Biology (Clifton, N.J.)* 2043 (2020), p. 1-12, ISSN : 1940-6029, DOI : 10.1007/978-1-4939-9698-8_1 (cf. p. 99).

[Ara+11] Bruno Aranda et al., « PSICQUIC and PSISCORE : accessing and scoring molecular interactions », eng, in : *Nature Methods* 8.*7* (juin 2011), p. 528-529, ISSN : 1548-7105, DOI : 10.1038/nmeth.1637 (cf. p. 40, 183, 184).

[Arb15] Brian S. Arbogast, « Phylogeography : The History and Formation of Species », in : *American Zoologist* 41.*1* (août 2015), p. 134-135, ISSN : 0003-1569, DOI : 10.1093/icb/41.1.134, eprint : https://academic.oup.com/icb/article-pdf/41/1/134/228494/i0003-1569-041-01-0134.pdf, URL : https://doi.org/10.1093/icb/41.1.134 (cf. p. 56).

[Ari+22] E. Ari et al., « A single early introduction governed viral diversity in the second wave of SARS-CoV-2 epidemic in Hungary », in : *Virus Evol* 8.*2* (juill. 2022), veac069 (cf. p. 56).

[Ash+19] Haim Ashkenazy et al., « Multiple Sequence Alignment Averaging Improves Phylogeny Reconstruction », en, in : *Systematic Biology* 68.*1* (jan. 2019), p. 117-130, ISSN : 1063-5157, DOI : 10.1093/sysbio/syy036, URL : https://academic.oup.com/sysbio/article/68/1/117/4996308 (visité le 08/11/2019) (cf. p. 140).

[Att+96] T. K. Attwood et al., « Progress With the PRINTS Protein Fingerprint Database », in : *Nucleic Acids Research* 24.*1* (jan. 1996), p. 182-188, ISSN : 0305-1048, DOI : 10.1093/nar/24.1.182, eprint : https://academic.oup.com/nar/article-pdf/24/1/182/6964079/24-1-182.pdf, URL : https://doi.org/10.1093/nar/24.1.182 (cf. p. 70).

[AYM13] Layal Al Ait, Zaher Yamak et Burkhard Morgenstern, « DIALIGN at GOBICS—multiple sequence alignment using various sources of external information », en, in : *Nucleic Acids Research* 41.*W1* (juill. 2013), Publisher :

Oxford Academic, W3-W7, ISSN : 0305-1048, DOI : 10.1093/nar/gkt283, URL : https://academic.oup.com/nar/article/41/W1/W3/1091179 (visité le 22/06/2020) (cf. p. 140, 141).

[BA99] A. Bairoch et R. Apweiler, « The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999 », in : *Nucleic Acids Res* 27.*1* (jan. 1999), p. 49-54 (cf. p. 30, 85).

[Ban+15] M. S. Bansal et al., « Improved gene tree error correction in the presence of horizontal gene transfer », in : *Bioinformatics* 31.*8* (avr. 2015), p. 1211-1218 (cf. p. 62).

[Bat+15] A. Bateman et al., « UniProt : a hub for protein information », in : *Nucleic Acids Res* 43.*Database issue* (jan. 2015), p. D204-212 (cf. p. 30).

[BB05] L. C. Bridges et R. D. Bowditch, « ADAM-Integrin Interactions : potential integrin regulated ectodomain shedding activity », in : *Curr Pharm Des* 11.*7* (2005), p. 837-847 (cf. p. 92).

[Bek+16] Mourad Bekhouche et al., « Determination of the substrate repertoire of ADAMTS2, 3 and 14 significantly broadens their functions and identifies extracellular matrix organization and TGF-beta signaling as primary targets. », en, in : *FASEB Journal* 5 (jan. 2016), Publisher : Federation of American Society for Experimental Biology, ISSN : 0892-6638, DOI : 10.1096/fj.15-279869, URL : https://orbi.uliege.be/handle/2268/195176 (visité le 10/01/2022) (cf. p. 184, 218).

[Ber+14] S. Beretta et al., « Modeling alternative splicing variants from RNA-Seq data with isoform graphs », in : *J Comput Biol* 21.*1* (jan. 2014), p. 16-40 (cf. p. 122).

[Ber05] J. Bergsten, « A review of long-branch attraction », in : *Cladistics* 21.*2* (avr. 2005), p. 163-193 (cf. p. 143).

[Bic+15] C. Bickelmann et al., « The molecular origin and evolution of dim-light vision in mammals », in : *Evolution* 69.*11* (nov. 2015), p. 2995-3003 (cf. p. 56).

[Blu+20] Matthias Blum et al., « The InterPro protein families and domains database : 20 years on », in : *Nucleic Acids Research* 49.*D1* (nov. 2020), p. D344-D354, ISSN : 0305-1048, DOI : 10.1093/nar/gkaa977, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7778928/ (visité le 05/01/2022) (cf. p. 67).

[BOD13] J. M. Beaulieu, B. C. O'Meara et M. J. Donoghue, « Identifying hidden rate changes in the evolution of a binary morphological character : the evolution of plant habit in campanulid angiosperms », in : *Syst Biol* 62.*5* (sept. 2013), p. 725-737 (cf. p. 56).

[Bor06] P. Bornstein, « EXTRACELLULAR MATRIX | Matricellular Proteins », in : *Encyclopedia of Respiratory Medicine*, sous la dir. de Geoffrey J. Laurent et Steven D. Shapiro, Oxford : Academic Press, 2006, p. 175-183, ISBN : 978-0-12-370879-3, DOI : https://doi.org/10.1016/B0-12-370879-6/00148-4, URL : https://www.sciencedirect.com/science/article/pii/B0123708796001484 (cf. p. 58).

[Bou+16] E. Boutet et al., « UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase : How to Use the Entry View », in : *Methods Mol Biol* 1374 (2016), p. 23-54 (cf. p. 65).

[Bou+19] R. Bouckaert et al., « BEAST 2.5 : An advanced software platform for Bayesian evolutionary analysis », in : *PLoS Comput Biol* 15.*4* (avr. 2019), e1006650 (cf. p. 55).

[Bru+15] Frédéric G Brunet et al., « The evolutionary conservation of the A Disintegrin-like and Metalloproteinase domain with Thrombospondin-1 motif metzincins across vertebrate species and their expression in teleost zebrafish », en, in : *BMC Evolutionary Biology* 15.*1* (déc. 2015), p. 22, ISSN : 1471-2148, DOI : 10.1186/s12862-015-0281-9, URL : https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-015-0281-9 (visité le 05/04/2021) (cf. p. 82, 83, 108, 109, 130).

[Bur+21] S. K. Burley et al., « RCSB Protein Data Bank : powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences », in : *Nucleic Acids Res* 49.*D1* (jan. 2021), p. D437-D451 (cf. p. 34).

[Bus+16] F. Busch et al., « Ancestral Tryptophan Synthase Reveals Functional Sophistication of Primordial Enzyme Complexes », in : *Cell Chem Biol* 23.*6* (juin 2016), p. 709-715 (cf. p. 56).

[BW21a] M. J. Binder et A. C. Ward, « The Role of the Metzincin Superfamily in Prostate Cancer Progression : A Systematic-Like Review », in : *Int J Mol Sci* 22.*7* (mars 2021) (cf. p. 79).

[BW21b] Marley J. Binder et Alister C. Ward, « The Role of the Metzincin Superfamily in Prostate Cancer Progression : A Systematic-Like Review », in : *International Journal of Molecular Sciences* 22.*7* (mars 2021), p. 3608, ISSN : 1422-0067, DOI : 10.3390/ijms22073608, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8036576/ (visité le 29/09/2021) (cf. p. 89).

[BZL11] A. G. Beristain, H. Zhu et P. C. Leung, « Regulated expression of ADAMTS-12 in human trophoblastic cells : a role for ADAMTS-12 in epithelial cell invasion ? », in : *PLoS One* 6.*4* (avr. 2011), e18473 (cf. p. 92).

[Cai+16] S. A. Cain et al., « ADAMTS-10 and -6 differentially regulate cell-cell junctions and focal adhesions », in : *Sci Rep* 6 (oct. 2016), p. 35956 (cf. p. 79).

[Cas00] J. Castresana, « Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis », in : *Mol Biol Evol* 17.*4* (avr. 2000), p. 540-552 (cf. p. 54, 128).

[CFS10] D. V. Cancherini, G. S. França et S. J. de Souza, « The role of exon shuffling in shaping protein-protein interaction networks », in : *BMC Genomics* 11 Suppl 5 (déc. 2010), S11 (cf. p. 58, 233).

[Cha+16a] M. Chatzou et al., « Multiple sequence alignment modeling : methods and applications », in : *Brief Bioinform* 17.*6* (nov. 2016), p. 1009-1023 (cf. p. 53).

[Cha+16b] Maria Chatzou et al., « Multiple sequence alignment modeling : methods and applications », en, in : *Briefings in Bioinformatics* 17.*6* (nov. 2016), p. 1009-1023, ISSN : 1467-5463, DOI : 10.1093/bib/bbv099, URL : https://academic.oup.com/bib/article/17/6/1009/2606431 (visité le 17/10/2019) (cf. p. 47).

[Che+22] M. Y. Chen et al., « Phylogenomics Uncovers Evolutionary Trajectory of Nitrogen Fixation in Cyanobacteria », in : *Mol Biol Evol* 39.*9* (sept. 2022) (cf. p. 56).

[CHW17] C. Chen, H. Huang et C. H. Wu, « Protein Bioinformatics Databases and Resources », in : *Methods Mol Biol* 1558 (2017), p. 3-39 (cf. p. 29).

[CK05] François Coste et Goulven Kerbellec, « A Similar Fragments Merging Approach to Learn Automata on Proteins », en, in : *Machine Learning : ECML 2005*, sous la dir. de João Gama et al., Lecture Notes in Computer Science, Berlin, Heidelberg : Springer, 2005, p. 522-529, isbn : 978-3-540-31692-3, doi : 10.1007/11564096_50 (cf. p. 122).

[CL15] S. Cal et C. López-Otín, « ADAMTS proteases and cancer », in : *Matrix Biol* 44-46 (2015), p. 77-85 (cf. p. 89).

[Col+04] L. A. Collins-Racie et al., « ADAMTS-8 exhibits aggrecanase activity and is expressed in human articular cartilage », in : *Matrix Biol* 23.*4* (juill. 2004), p. 219-230 (cf. p. 92).

[Col+19] Alain Colige et al., « Proteomic discovery of substrates of the cardiovascular protease ADAMTS7 », in : *The Journal of Biological Chemistry* 294.*20* (mai 2019), p. 8037-8045, issn : 0021-9258, doi : 10.1074/jbc.RA119.007492, url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6527163/ (visité le 05/01/2022) (cf. p. 184).

[Cos22] François Coste, *Protomata 2.0*, http://protomata-learner.genouest.org, 2022 (cf. p. 72, 150).

[Cro+14] Graham Cromar et al., « New Tricks for "Old" Domains : How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM », in : *Genome Biology and Evolution* 6.*10* (oct. 2014), p. 2897-2917, issn : 1759-6653, doi : 10.1093/gbe/evu228, url : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224354/ (visité le 27/11/2019) (cf. p. 223, 234).

[CSG09] Salvador Capella-Gutiérrez, José M. Silla-Martínez et Toni Gabaldón, « trimAl : a tool for automated alignment trimming in large-scale phylogenetic analyses », in : *Bioinformatics* 25.*15* (août 2009), p. 1972-1973, issn : 1367-4803, doi : 10.1093/bioinformatics/btp348, url : https://doi.org/10.1093/bioinformatics/btp348 (visité le 01/04/2021) (cf. p. 54, 128, 141).

[DB07] K. Dolinski et D. Botstein, « Orthology and functional conservation in eukaryotes », in : *Annu Rev Genet* 41 (2007), p. 465-507 (cf. p. 190).

[DC12] L. Dib et A. Carbone, « Protein fragments : functional and structural roles of their coevolution networks », in : *PLoS One* 7.*11* (2012), e48124 (cf. p. 234).

[DC15] K. Djinovic-Carugo et O. Carugo, « Missing strings of residues in protein crystal structures », in : *Intrinsically Disord Proteins* 3.*1* (2015), e1095697 (cf. p. 34).

[Dem+10] Ilya V. Demidyuk et al., « Propeptides as modulators of functional activity of proteases », eng, in : *Biomolecular Concepts* 1.*3-4* (oct. 2010), p. 305-322, ISSN : 1868-5021, DOI : 10.1515/bmc.2010.025 (cf. p. 216, 235).

[Den+02] Minghua Deng et al., « Inferring Domain–Domain Interactions From Protein–Protein Interactions », in : *Genome Research* 12.*10* (oct. 2002), p. 1540-1548, ISSN : 1088-9051, DOI : 10.1101/gr.153002, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC187530/ (visité le 01/04/2021) (cf. p. 38).

[Dic+03] Sally C. Dickinson et al., « Cleavage of cartilage oligomeric matrix protein (thrombospondin-5) by matrix metalloproteinases and a disintegrin and metalloproteinase with thrombospondin motifs », eng, in : *Matrix Biology : Journal of the International Society for Matrix Biology* 22.*3* (mai 2003), p. 267-278, ISSN : 0945-053X, DOI : 10.1016/s0945-053x(03)00034-9 (cf. p. 218, 224).

[DM15] G. Dey et T. Meyer, « Phylogenetic Profiling for Probing the Modular Architecture of the Human Genome », in : *Cell Syst* 1.*2* (août 2015), p. 106-115 (cf. p. 75).

[Doh+20] Elias Dohmen et al., « The modular nature of protein evolution : domain rearrangement rates across eukaryotic life », in : *BMC Evolutionary Biology* 20.*1* (fév. 2020), p. 30, ISSN : 1471-2148, DOI : 10.1186/s12862-020-1591-0, URL : https://doi.org/10.1186/s12862-020-1591-0 (visité le 24/03/2020) (cf. p. 36).

[Doo95] R. F. Doolittle, « The multiplicity of domains in proteins », in : *Annu Rev Biochem* 64 (1995), p. 287-314 (cf. p. 36).

[Doy+11] J. P. Doyon et al., « Models, algorithms and programs for phylogeny reconciliation », in : *Brief Bioinform* 12.*5* (sept. 2011), p. 392-400 (cf. p. 60).

[Duc+18] W. Duchemin et al., « RecPhyloXML : a format for reconciled gene trees », in : *Bioinformatics* 34.*21* (nov. 2018), p. 3646-3652 (cf. p. 63).

[Dud+17] G. Dudas et al., « Virus genomes reveal factors that spread and sustained the Ebola epidemic », in : *Nature* 544.*7650* (avr. 2017), p. 309-315 (cf. p. 56).

[ED09] P. K. Endress et J. A. Doyle, « Reconstructing the ancestral angiosperm flower and its initial specializations », in : *Am J Bot* 96.*1* (jan. 2009), p. 22-66 (cf. p. 56).

[Edg04] Robert C. Edgar, « MUSCLE : multiple sequence alignment with high accuracy and high throughput », eng, in : *Nucleic Acids Research* 32.*5* (2004), p. 1792-1797, ISSN : 1362-4962, DOI : 10.1093/nar/gkh340 (cf. p. 140, 141).

[Edw+11] C. J. Edwards et al., « Ancient hybridization and an Irish origin for the modern polar bear matriline », in : *Curr Biol* 21.*15* (août 2011), p. 1251-1258 (cf. p. 56).

[Eis98] Jonathan Eisen, « Phylogenomics : Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis », in : *Genome research* 8 (mars 1998), p. 163-7, DOI : 10.1101/gr.8.3.163 (cf. p. 76, 201).

[EK15] David M. Emms et Steven Kelly, « OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy », in : *Genome Biology* 16.*1* (2015), ISSN : 1474-7596, DOI : 10.1186/s13059-015-0721-2, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531804/ (visité le 27/01/2020) (cf. p. 105, 114, 117).

[EK19] David M. Emms et Steven Kelly, « OrthoFinder : phylogenetic orthology inference for comparative genomics », in : *Genome Biology* 20.*1* (nov. 2019), p. 238, ISSN : 1474-760X, DOI : 10.1186/s13059-019-1832-y, URL : https://doi.org/10.1186/s13059-019-1832-y (visité le 23/01/2020) (cf. p. 105, 114, 117, 125).

[ELI] ELIXIR, *InterPro Online Documentation*, URL : https://www.ebi.ac.uk/interpro/about/interpro/ (visité le 18/10/2022) (cf. p. 68).

[EM16] Patrick A. Eyers et James M. Murphy, « The evolving world of pseudoenzymes : proteins, prejudice and zombies », in : *BMC Biology* 14 (nov. 2016), ISSN : 1741-7007, DOI : 10.1186/s12915-016-0322-x, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5106787/ (visité le 28/10/2019) (cf. p. 39).

[Eng+05] Barbara E. Engelhardt et al., « Protein Molecular Function Prediction by Bayesian Phylogenomics », en, in : *PLOS Computational Biology* 1.5 (oct. 2005), Publisher : Public Library of Science, e45, ISSN : 1553-7358, DOI : 10.1371/journal.pcbi.0010045, URL : https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0010045 (visité le 15/09/2020) (cf. p. 76, 201).

[Far+14] C. M. Farrell et al., « Current status and new features of the Consensus Coding Sequence database », in : *Nucleic Acids Res* 42.*Database issue* (jan. 2014), p. D865-872 (cf. p. 29).

[Fel85] Joseph Felsenstein, « Confidence Limits on Phylogenies : An Approach Using the Bootstrap », in : *Evolution* 39.*4* (1985), Publisher : [Society for the Study of Evolution, Wiley], p. 783-791, ISSN : 0014-3820, DOI : 10.2307/2408678, URL : https://www.jstor.org/stable/2408678 (visité le 01/04/2021) (cf. p. 54).

[FG13] Hai Fang et Julian Gough, « dcGO : database of domain-centric ontologies on functions, phenotypes, diseases and more », in : *Nucleic Acids Research* 41.*Database issue* (jan. 2013), p. D536-D544, ISSN : 0305-1048, DOI : 10.1093/nar/gks1080, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531119/ (visité le 20/10/2020) (cf. p. 201, 211).

[Fin10] Audrey Finkler, « Modèle d'évolution avec dépendance au contexte et Corrections de statistiques d'adéquation en présence de zéros aléatoires », Theses, Université de Strasbourg, juin 2010, URL : https://tel.archives-ouvertes.fr/tel-00490844 (cf. p. 51).

[Fon+21] Tania Fontanil et al., « Hyalectanase Activities by the ADAMTS Metalloproteases », eng, in : *International Journal of Molecular Sciences* 22.*6* (mars 2021), p. 2988, ISSN : 1422-0067, DOI : 10.3390/ijms22062988 (cf. p. 184, 224).

[Fos04] P. G. Foster, « Modeling compositional heterogeneity », in : *Syst Biol* 53.*3* (juin 2004), p. 485-495 (cf. p. 55).

[Fou+14] Simon J. Foulcer et al., « Determinants of versican-V1 proteoglycan processing by the metalloproteinase ADAMTS5 », eng, in : *The Journal of Biological Chemistry* 289.*40* (oct. 2014), p. 27859-27873, ISSN : 1083-351X, DOI : 10.1074/jbc.M114.573287 (cf. p. 224).

[Fri06] I. Friedberg, « Automated protein function prediction–the genomic challenge », in : *Brief Bioinform* 7.*3* (sept. 2006), p. 225-242 (cf. p. 67).

[Gau+11] Pascale Gaudet et al., « Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium », in : *Briefings in Bioinformatics* 12.*5* (sept. 2011), p. 449-462, ISSN : 1467-5463, DOI : 10.1093/bib/bbr042, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178059/ (visité le 20/11/2019) (cf. p. 76, 201, 211).

[GBB22] N. Guillaudeux, C. Belleannée et S. Blanquart, « Identifying genes with conserved splicing structure and orthologous isoforms in human, mouse and dog », in : *BMC Genomics* 23.*1* (mars 2022), p. 216 (cf. p. 122).

[Gen+07] Christi Gendron et al., « Proteolytic activities of human ADAMTS-5 : comparative studies with ADAMTS-4 », eng, in : *The Journal of Biological Chemistry* 282.*25* (juin 2007), p. 18294-18306, ISSN : 0021-9258, DOI : 10.1074/jbc.M701523200 (cf. p. 92, 224, 228).

[GH18] J. L. Geoghegan et E. C. Holmes, « The phylogenomics of evolving virus virulence », in : *Nat Rev Genet* 19.*12* (déc. 2018), p. 756-769 (cf. p. 56).

[GK13] T. Gabaldón et E. V. Koonin, « Functional and evolutionary implications of gene orthology », in : *Nat Rev Genet* 14.*5* (mai 2013), p. 360-366 (cf. p. 190).

[GKT09] Bastien D. Gomperts, IJsbrand M. Kramer et Peter E.R. Tatham, « Chapter 24 - Protein Domains and Signal Transduction », in : *Signal Transduction (Second Edition)*, sous la dir. de Bastien D. Gomperts, IJsbrand M. Kramer et Peter E.R. Tatham, Second Edition, San Diego : Academic Press, 2009, p. 763-790, ISBN : 978-0-12-369441-6, DOI : https://doi.org/10.1016/B978-0-12-369441-6.00024-6, URL : https://

www.sciencedirect.com/science/article/pii/B9780123694416000246 (cf. p. 58).

[Gla+05] Sonya S. Glasson et al., « Deletion of active ADAMTS5 prevents cartilage degradation in a murine model of osteoarthritis », eng, in : *Nature* 434.*7033* (mars 2005), p. 644-648, ISSN : 1476-4687, DOI : 10.1038/nature03369 (cf. p. 89, 92, 224, 228).

[GLC15] R. de Groot, D. A. Lane et J. T. Crawley, « The role of the ADAMTS13 cysteine-rich domain in VWF binding and proteolysis », in : *Blood* 125.*12* (mars 2015), p. 1968-1975 (cf. p. 92).

[GME87] M. Gribskov, A. D. McLachlan et D. Eisenberg, « Profile analysis : detection of distantly related proteins », in : *Proc Natl Acad Sci U S A* 84.*13* (juill. 1987), p. 4355-4358 (cf. p. 70).

[Goo+79] Morris Goodman et al., « Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences », in : *Systematic Zoology* 28.*2* (1979), p. 132-163, ISSN : 00397989, URL : http://www.jstor.org/stable/2412519 (visité le 12/10/2022) (cf. p. 60).

[Gro+09] R. de Groot et al., « Essential role of the disintegrin-like domain in ADAMTS13 function », in : *Blood* 113.*22* (mai 2009), p. 5609-5616 (cf. p. 92).

[Gui+10] Stéphane Guindon et al., « New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0 », en, in : *Systematic Biology* 59.*3* (mars 2010), p. 307-321, ISSN : 1076-836X, 1063-5157, DOI : 10.1093/sysbio/syq010, URL : https://academic.oup.com/sysbio/article/59/3/307/1702850 (visité le 01/04/2021) (cf. p. 55, 128, 141, 160).

[HA15] D. Hubmacher et S. S. Apte, « ADAMTS proteins as modulators of microfibril formation and function », in : *Matrix Biol* 47 (sept. 2015), p. 34-43 (cf. p. 79).

[Han+20] X. Han et al., « Ab Initio Construction and Evolutionary Analysis of Protein-Coding Gene Families with Partially Homologous Relationships : Closely Related Drosophila Genomes as a Case Study », in : *Genome Biol Evol* 12.*3* (mars 2020), p. 185-202 (cf. p. 58).

[HB01] J. P. Huelsenbeck et J. P. Bollback, « Empirical and hierarchical Bayesian estimation of ancestral states », in : *Syst Biol* 50.*3* (juin 2001), p. 351-366 (cf. p. 56).

[HLM21] Edwin Rodríguez Horta, Alejandro Lage-Castellanos et Roberto Mulet, « Ancestral Sequence Reconstruction for Co-evolutionary models », in : *arXiv* (2021), doi : 10.48550/arxiv.2108.03801, url : https://app.dimensions.ai/details/publication/pub.1140317022 (cf. p. 56).

[Höh+16] S. Höhna et al., « RevBayes : Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language », in : *Syst Biol* 65.*4* (juill. 2016), p. 726-736 (cf. p. 55).

[HR01] J. P. Huelsenbeck et F. Ronquist, « MRBAYES : Bayesian inference of phylogenetic trees », in : *Bioinformatics* 17.*8* (août 2001), p. 754-755 (cf. p. 55).

[Hux+05] Julie Huxley-Jones et al., « The characterisation of six ADAMTS proteases in the basal chordate Ciona intestinalis provides new insights into the vertebrate ADAMTS family », en, in : *The International Journal of Biochemistry & Cell Biology* 37.*9* (sept. 2005), p. 1838-1845, issn : 13572725, doi : 10.1016/j.biocel.2005.03.009, url : https://linkinghub.elsevier.com/retrieve/pii/S1357272505001123 (visité le 05/04/2021) (cf. p. 82, 83, 109, 130).

[Hux+07] Julie Huxley-Jones et al., « The evolution of the vertebrate metzincins ; insights from Ciona intestinalis and Danio rerio », en, in : *BMC Evolutionary Biology* 7.*1* (2007), p. 63, issn : 14712148, doi : 10.1186/1471-2148-7-63, url : http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-7-63 (visité le 05/04/2021) (cf. p. 79, 80, 82, 83, 130, 142, 143).

[Hyn12] R. O. Hynes, « The evolution of metazoan extracellular matrix », in : *J Cell Biol* 196.*6* (mars 2012), p. 671-679 (cf. p. 108, 223).

[IG12] Hiroaki Iwata et Osamu Gotoh, « Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features », en, in : *Nucleic Acids Research* 40.*20* (nov. 2012), e161-e161, issn : 0305-1048, doi : 10.1093/nar/gks708, url : https://academic.oup.com/nar/article/40/20/e161/2414522 (visité le 09/03/2020) (cf. p. 121).

[Ish+19] Sohta A Ishikawa et al., « A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios », in : *Molecular Biology and Evolution* 36.*9* (sept. 2019), p. 2069-2085, ISSN : 0737-4038, DOI : 10.1093/molbev/msz131, URL : https://doi.org/10.1093/molbev/msz131 (visité le 01/04/2021) (cf. p. 56, 188, 196).

[Jef19] Constance J. Jeffery, « The demise of catalysis, but new functions arise : pseudoenzymes as the phoenixes of the protein world », en, in : *Biochemical Society Transactions* 47.*1* (fév. 2019), p. 371-379, ISSN : 0300-5127, DOI : 10.1042/BST20180473, URL : /biochemsoctrans/article/47/1/371/140/The-demise-of-catalysis-but-new-functions-arise (visité le 28/10/2019) (cf. p. 45).

[Jia+21] L. Jiang et al., « ADAMTS5 in Osteoarthritis : Biological Functions, Regulatory Network, and Potential Targeting Therapies », in : *Front Mol Biosci* 8 (2021), p. 703110 (cf. p. 229, 235).

[JLS17] Jesper Jansson, Zhaoxian Li et Wing-Kin Sung, « On finding the Adams consensus tree », en, in : *Information and Computation* 256 (oct. 2017), p. 334-347, ISSN : 0890-5401, DOI : 10.1016/j.ic.2017.08.002, URL : http://www.sciencedirect.com/science/article/pii/S0890540117301402 (visité le 19/08/2020) (cf. p. 135).

[Jum+21] J. Jumper et al., « Highly accurate protein structure prediction with AlphaFold », in : *Nature* 596.*7873* (août 2021), p. 583-589 (cf. p. 34, 85-88).

[Kal+19] K. Kalogeropoulos et al., « Protease Activity Profiling of Snake Venoms Using High-Throughput Peptide Screening », in : *Toxins (Basel)* 11.*3* (mars 2019) (cf. p. 79).

[Kar+21] N. K. Karamanos et al., « A guide to the composition and functions of the extracellular matrix », in : *FEBS J* 288.*24* (déc. 2021), p. 6850-6912 (cf. p. 223).

[Kas+04] Masahide Kashiwagi et al., « Altered proteolytic activities of ADAMTS-4 expressed by C-terminal processing », eng, in : *The Journal of Biological Chemistry* 279.*11* (mars 2004), p. 10109-10119, ISSN : 0021-9258, DOI : 10.1074/jbc.M312123200 (cf. p. 224).

[Kat+02] K. Katoh et al., « MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform », in : *Nucleic Acids Res* 30.*14* (juill. 2002), p. 3059-3066 (cf. p. 140, 141).

[Kaw+09] T. Kawashima et al., « Domain shuffling and the evolution of vertebrates », in : *Genome Res* 19.*8* (août 2009), p. 1393-1403 (cf. p. 58, 233).

[Kec93] John D. Kececioglu, « The Maximum Weight Trace Problem in Multiple Sequence Alignment », in : *Combinatorial Pattern Matching, 4th Annual Symposium, CPM 93, Padova, Italy, June 2-4, 1993, Proceedings*, sous la dir. d'Alberto Apostolico et al., t. 684, Lecture Notes in Computer Science, Springer, 1993, p. 106-119, DOI : 10.1007/BFb0029800, URL : https://doi.org/10.1007/BFb0029800 (cf. p. 73).

[Kel+15] Richard Kelwick et al., « The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family », eng, in : *Genome Biology* 16 (mai 2015), p. 113, ISSN : 1474-760X, DOI : 10.1186/s13059-015-0676-3 (cf. p. 79, 82, 184, 216).

[Ker08] Goulven Kerbellec, « Apprentissage d'automates modélisant des familles de séquences protéiques. (Learning automata modelling families of protein sequences) », PhD Thesis, University of Rennes 1, France, 2008, URL : https://tel.archives-ouvertes.fr/tel-00327938 (cf. p. 72, 178, 238).

[Kod+15] Y. Kodama et al., « The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data », in : *Nucleic Acids Res* 43.*Database issue* (jan. 2015), p. 18-22 (cf. p. 29).

[Kol+21] R. Kolodny et al., « Bridging Themes : Short Protein Segments Found in Different Architectures », in : *Mol Biol Evol* 38.*6* (mai 2021), p. 2191-2208 (cf. p. 36).

[Kol21] R. Kolodny, « Searching protein space for ancient sub-domain segments », in : *Curr Opin Struct Biol* 68 (juin 2021), p. 105-112 (cf. p. 36).

[Koz+19] A. M. Kozlov et al., « RAxML-NG : a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference », in : *Bioinformatics* 35.*21* (nov. 2019), p. 4453-4455 (cf. p. 55).

[Kro+94] A. Krogh et al., « Hidden Markov models in computational biology. Applications to protein modeling », in : *J Mol Biol* 235.*5* (fév. 1994), p. 1501-1531 (cf. p. 69).

[Kul+07] T. Kulikova et al., « EMBL Nucleotide Sequence Database in 2006 », in : *Nucleic Acids Res* 35.*Database issue* (jan. 2007), p. 16-20 (cf. p. 29).

[KWK02] E. V. Koonin, Y. I. Wolf et G. P. Karev, « The structure of the protein universe and genome evolution », in : *Nature* 420.*6912* (nov. 2002), p. 218-223 (cf. p. 36).

[Kwu+22] Min Jung Kwun et al., « Post-vaccine epidemiology of serotype 3 pneumococci identifies transformation inhibition through prophage-driven alteration of a non-coding RNA », in : *bioRxiv* (2022), DOI : 10.1101/2022.09.21.508813, eprint : https://www.biorxiv.org/content/early/2022/09/21/2022.09.21.508813.full.pdf, URL : https://www.biorxiv.org/content/early/2022/09/21/2022.09.21.508813 (cf. p. 56).

[KYT20] P. Kapli, Z. Yang et M. J. Telford, « Phylogenetic tree building in the genomic age », in : *Nat Rev Genet* 21.*7* (juill. 2020), p. 428-444 (cf. p. 55).

[Lac+08] Vincent Lacroix et al., « Exact Transcriptome Reconstruction from Short Sequence Reads », in : *Algorithms in Bioinformatics*, sous la dir. de Keith A. Crandall et Jens Lagergren, Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 50-63, ISBN : 978-3-540-87361-7 (cf. p. 122).

[Lau+21] L. Laursen et al., « Divergent Evolution of a Protein-Protein Interaction Revealed through Ancestral Sequence Reconstruction and Resurrection », in : *Mol Biol Evol* 38.*1* (jan. 2021), p. 152-167 (cf. p. 201, 211).

[LB18] Lei Li et Mukul S. Bansal, « An Integer Linear Programming Solution for the Domain-Gene-Species Reconciliation Problem », in : *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, event-place : Washington, DC, USA, New York, NY, USA : ACM, 2018, p. 386-397, ISBN : 978-1-4503-5794-4, DOI : 10.1145/3233547.3233603, URL : http://doi.acm.org/10.1145/3233547.3233603 (visité le 17/10/2019) (cf. p. 61).

[LB19a] Ivica Letunic et Peer Bork, « Interactive Tree Of Life (iTOL) v4 : recent updates and new developments », in : *Nucleic Acids Research* 47.*W1* (juill. 2019), W256-W259, ISSN : 0305-1048, DOI : 10.1093/nar/gkz239, URL : https://doi.org/10.1093/nar/gkz239 (visité le 01/04/2021) (cf. p. 109, 153, 207).

[LB19b] L. Li et M. S. Bansal, « An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution », in : *IEEE/ACM Trans Comput Biol Bioinform* 16.*1* (2019), p. 63-76 (cf. p. 58, 61, 157, 159, 233).

[LB19c] Lei Li et Mukul Bansal, « Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model », in : jan. 2019, p. 73-86, ISBN : 978-3-030-20241-5, DOI : 10.1007/978-3-030-20242-2_7 (cf. p. 62, 63, 157, 162).

[LC11] Carine Le Goff et Valérie Cormier-Daire, « The ADAMTS(L) family and human genetic disorders », in : *Human Molecular Genetics* 20.*R2* (oct. 2011), R163-R167, ISSN : 0964-6906, DOI : 10.1093/hmg/ddr361, URL : https://doi.org/10.1093/hmg/ddr361 (visité le 18/03/2022) (cf. p. 89).

[Le +08] C. Le Goff et al., « ADAMTSL2 mutations in geleophysic dysplasia demonstrate a role for ADAMTS-like proteins in TGF-beta bioavailability regulation », in : *Nat Genet* 40.*9* (sept. 2008), p. 1119-1123 (cf. p. 80).

[Le +11] C. Le Goff et al., « Mutations in the TGFβ binding-protein-like domain 5 of FBN1 are responsible for acromicric and geleophysic dysplasias », in : *Am J Hum Genet* 89.*1* (juill. 2011), p. 7-14 (cf. p. 80).

[Led+21] Cédric Leduc et al., « In vivo N-Terminomics Highlights Novel Functions of ADAMTS2 and ADAMTS14 in Skin Collagen Matrix Building », en, in : *Frontiers in Molecular Biosciences* 8 (mars 2021), p. 643178, ISSN : 2296-889X, DOI : 10.3389/fmolb.2021.643178, URL : https://www.frontiersin.org/articles/10.3389/fmolb.2021.643178/full (visité le 31/05/2021) (cf. p. 184).

[Liu+06a] Chuan-Ju Liu et al., « ADAMTS-7 : a metalloproteinase that directly binds to and degrades cartilage oligomeric matrix protein », eng, in : *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 20.*7* (mai 2006), p. 988-990, ISSN : 1530-6860, DOI : 10.1096/fj.05-3877fje (cf. p. 218, 220, 223).

[Liu+06b] Chuan-ju Liu et al., « ADAMTS-12 Associates with and Degrades Cartilage Oligomeric Matrix Protein », in : *The Journal of biological chemistry* 281.*23* (juin 2006), p. 15800-15808, ISSN : 0021-9258, DOI : 10.1074/jbc.M513433200, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1483932/ (visité le 10/01/2022) (cf. p. 218, 220, 223).

[Liu09] C. J. Liu, « The role of ADAMTS-7 and ADAMTS-12 in the pathogenesis of arthritis », in : *Nat Clin Pract Rheumatol* 5.*1* (jan. 2009), p. 38-45 (cf. p. 218).

[LLB09] N. Lartillot, T. Lepage et S. Blanquart, « PhyloBayes 3 : a Bayesian software package for phylogenetic reconstruction and molecular dating », in : *Bioinformatics* 25.*17* (sept. 2009), p. 2286-2288 (cf. p. 55).

[Löy14] A. Löytynoja, « Phylogeny-aware alignment with PRANK », in : *Methods Mol Biol* 1079 (2014), p. 155-170 (cf. p. 140, 141).

[MA18] Timothy J. Mead et Suneel S. Apte, « ADAMTS proteins in human disorders », in : *Matrix biology : journal of the International Society for Matrix Biology* 71-72 (oct. 2018), p. 225-239, ISSN : 0945-053X, DOI : 10.1016/j.matbio.2018.06.002, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6146047/ (visité le 17/10/2019) (cf. p. 79, 84, 89).

[Mad97] Wayne P. Maddison, « Gene Trees in Species Trees », in : *Systematic Biology* 46.*3* (sept. 1997), p. 523-536, ISSN : 1063-5157, DOI : 10.1093/sysbio/46.3.523, eprint : https://academic.oup.com/sysbio/article-pdf/46/3/523/19501929/46-3-523.pdf, URL : https://doi.org/10.1093/sysbio/46.3.523 (cf. p. 60).

[Mao+17] R. Maor et al., « Temporal niche expansion in mammals from a nocturnal ancestor after dinosaur extinction », in : *Nat Ecol Evol* 1.*12* (déc. 2017), p. 1889-1895 (cf. p. 56).

[Mar+06] R. L. Marsden et al., « Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space », in : *Nucleic Acids Res* 34.*3* (2006), p. 1066-1080 (cf. p. 36).

[Mar+12] B. Marazzi et al., « Locating evolutionary precursors on a phylogenetic tree », in : *Evolution* 66.*12* (déc. 2012), p. 3918-3930 (cf. p. 56).

[Mar+95] A. Marchese et al., « Cloning and chromosomal mapping of three novel genes, GPR9, GPR10, and GPR14, encoding receptors related to interleukin 8, neuropeptide Y, and somatostatin receptors », in : *Genomics* 29.*2* (sept. 1995), p. 335-344 (cf. p. 122).

[Mas22] M. L. Mascotti, « Resurrecting Enzymes by Ancestral Sequence Reconstruction », in : *Methods Mol Biol* 2397 (2022), p. 111-136 (cf. p. 56).

[Md +15] A. S. Md Mukarram Hossain et al., « Evidence of Statistical Inconsistency of Phylogenetic Methods in the Presence of Multiple Sequence Alignment Uncertainty », in : *Genome Biol Evol* 7.*8* (juill. 2015), p. 2102-2116 (cf. p. 140).

[MDT21] Hugo Menet, Vincent Daubin et Eric Tannier, « Phylogenetic reconciliation », This preprint was submitted to Plos Comp Biol Topic Pages, thus its style and format are intented to fit a wikipedia article., juin 2021, URL : https://hal.archives-ouvertes.fr/hal-03258402 (cf. p. 60).

[MFE17] James M. Murphy, Hesso Farhan et Patrick A. Eyers, « Bio-Zombie : the rise of pseudoenzymes in biology », en, in : *Biochemical Society Transactions* 45.*2* (avr. 2017), p. 537-544, ISSN : 0300-5127, DOI : 10.1042/BST20160400, URL : /biochemsoctrans/article/45/2/537/67103/Bio-Zombie-the-rise-of-pseudoenzymes-in-biology (visité le 28/10/2019) (cf. p. 39, 130).

[Min+20] B. Q. Minh et al., « IQ-TREE 2 : New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era », in : *Mol Biol Evol* 37.*5* (mai 2020), p. 1530-1534 (cf. p. 55).

[Mir+15] Siavash Mirarab et al., « PASTA : Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences », in : *Journal of Computational Biology* 22.*5* (mai 2015), p. 377-386, ISSN : 1066-5277, DOI : 10.1089/cmb.2014.0156, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4424971/ (visité le 05/11/2019) (cf. p. 127).

[Mis+21] Jaina Mistry et al., « Pfam : The protein families database in 2021 », in : *Nucleic Acids Research* 49.*D1* (jan. 2021), p. D412-D419, ISSN : 0305-1048, DOI : 10.1093/nar/gkaa913, URL : https://doi.org/10.1093/nar/gkaa913 (visité le 01/04/2021) (cf. p. 69, 152).

[Mit+05] Laureane Mittaz et al., « Neonatal calyceal dilation and renal fibrosis resulting from loss of Adamts-1 in mouse kidney is due to a developmental dysgenesis », en, in : *Nephrology Dialysis Transplantation* 20.*2* (fév. 2005), p. 419-423, ISSN : 0931-0509, DOI : 10.1093/ndt/gfh603, URL : https://academic.oup.com/ndt/article/20/2/419/1835795 (visité le 25/11/2019) (cf. p. 89).

[ML08] Gabriel Moreno-Hagelsieb et Kristen Latimer, « Choosing BLAST options for better detection of orthologs as reciprocal best hits », en, in : *Bioinformatics* 24.*3* (fév. 2008), p. 319-324, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/btm585, URL : https://academic.oup.com/bioinformatics/article/24/3/319/252715 (visité le 22/11/2019) (cf. p. 108).

[Moh+21a] Y. Mohamedi et al., « ADAMTS-12 : Functions and Challenges for a Complex Metalloprotease », in : *Front Mol Biosci* 8 (2021), p. 686763 (cf. p. 233).

[Moh+21b] Yamina Mohamedi et al., « ADAMTS-12 : Functions and Challenges for a Complex Metalloprotease », in : *Frontiers in Molecular Biosciences* 8 (2021), p. 378, ISSN : 2296-889X, DOI : 10.3389/fmolb.2021.686763, URL : https://www.frontiersin.org/article/10.3389/fmolb.2021.686763 (visité le 05/01/2022) (cf. p. 184).

[Mor99] Burkhard Morgenstern, « DIALIGN 2 : improvement of the segment-to-segment approach to multiple sequence alignment », in : *Bioinform.* 15.*3* (1999), p. 211-218, DOI : 10.1093/bioinformatics/15.3.211, URL : https://doi.org/10.1093/bioinformatics/15.3.211 (cf. p. 73, 140, 141, 151).

[Müh+20] B. Mühlemann et al., « Diverse variola virus (smallpox) strains were widespread in northern Europe in the Viking Age », in : *Science* 369.*6502* (juill. 2020) (cf. p. 56).

[Nak09] Luay Nakhleh, *The Problem Solving Handbook for Computational Biology and Bioinformatics, chapter Evolutionary phylogenetic networks : models and issues*, 2009 (cf. p. 60).

[Nan+19] Sumeda Nandadasa et al., « Secreted metalloproteases ADAMTS9 and ADAMTS20 have a non-canonical role in ciliary vesicle growth during ciliogenesis », in : *Nature Communications* 10 (fév. 2019), p. 953, ISSN : 2041-1723, DOI : 10.1038/s41467-019-08520-7, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6393521/ (visité le 25/01/2022) (cf. p. 184).

[Nev+19] Yannis Nevers et al., « OrthoInspector 3.0 : open portal for comparative genomics », in : *Nucleic Acids Research* 47.*Database issue* (jan. 2019), p. D411-D418, ISSN : 0305-1048, DOI : 10.1093/nar/gky1068, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323921/ (visité le 13/11/2019) (cf. p. 108-110).

[NHH00] C. Notredame, D. G. Higgins et J. Heringa, « T-Coffee : A novel method for fast and accurate multiple sequence alignment », in : *J Mol Biol* 302.*1* (sept. 2000), p. 205-217 (cf. p. 140, 141).

[Nic+05] Ainsley C Nicholson et al., « Functional evolution of ADAMTS genes : Evidence from analyses of phylogeny and gene organization », en, in : *BMC Evolutionary Biology* 5.*1* (déc. 2005), p. 11, ISSN : 1471-2148, DOI : 10.1186/1471-2148-5-11, URL : https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-5-11 (visité le 05/04/2021) (cf. p. 82, 83, 130).

[NRI06] L Nakhleh, D Ruths et H Innan, « Meta-analysis and Combining Information in Genetics and Genomics », in : *Chapman & Hall/CRC. chapter Gene trees, species trees, and species networks* 1 (2006), p. 1-27 (cf. p. 60).

[NW70] S. B. Needleman et C. D. Wunsch, « A general method applicable to the search for similarities in the amino acid sequence of two proteins », in : *J Mol Biol* 48.*3* (mars 1970), p. 443-453 (cf. p. 102).

[Oak17] T. H. Oakley, « Furcation and fusion : The phylogenetics of evolutionary novelty », in : *Dev Biol* 431.*1* (nov. 2017), p. 69-76 (cf. p. 58, 157, 233).

[OR06] T Heath Ogden et Michael S Rosenberg, « Multiple Sequence Alignment Accuracy and Phylogenetic Inference », in : *Systematic Biology* 55.*2* (avr. 2006), p. 314-328, ISSN : 1063-5157, DOI : 10.1080/10635150500541730, eprint : https://academic.oup.com/sysbio/article-pdf/55/2/314/26557266/10635150500541730.pdf, URL : https://doi.org/10.1080/10635150500541730 (cf. p. 140).

[Pag99] Mark Pagel, « The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies », in : *Systematic biology* 48.*3* (1999), p. 612-622 (cf. p. 56).

[Pat21] L. Patthy, « Exon Shuffling Played a Decisive Role in the Evolution of the Genetic Toolkit for the Multicellular Body Plan of Metazoa », in : *Genes (Basel)* 12.*3* (mars 2021) (cf. p. 58).

[Pat96] L. Patthy, « Exon shuffling and other ways of module exchange », in : *Matrix Biol* 15.*5* (nov. 1996), p. 301-310 (cf. p. 58, 233).

[Paz22] F. Pazos, « Computational prediction of protein functional sites-Applications in biotechnology and biomedicine », in : *Adv Protein Chem Struct Biol* 130 (2022), p. 39-57 (cf. p. 47, 48).

[PDA10] M. N. Price, P. S. Dehal et A. P. Arkin, « FastTree 2–approximately maximum-likelihood trees for large alignments », in : *PLoS One* 5.*3* (mars 2010), e9490 (cf. p. 55).

[Pel+99] M. Pellegrini et al., « Assigning protein functions by comparative genome analysis : protein phylogenetic profiles », in : *Proc Natl Acad Sci U S A* 96.*8* (avr. 1999), p. 4285-4288 (cf. p. 75).

[Pen13] D. Penny, « Rewriting evolution–"been there, done that" », in : *Genome Biol Evol* 5.*5* (2013), p. 819-821 (cf. p. 54).

[Per+19] L. Perfetto et al., « CausalTAB : the PSI-MITAB 2.8 updated format for signalling data representation and dissemination », in : *Bioinformatics* 35.*19* (oct. 2019), p. 3779-3785 (cf. p. 40).

[Pér+20] Selene Pérez-García et al., « Profile of Matrix-Remodeling Proteinases in Osteoarthritis : Impact of Fibronectin », en, in : *Cells* 9.*1* (jan. 2020), p. 40, doi : 10.3390/cells9010040, url : https://www.mdpi.com/2073-4409/9/1/40 (visité le 06/01/2020) (cf. p. 89).

[Pey+11] F. Peysselon et al., « Intrinsic disorder of the extracellular matrix », in : *Mol Biosyst* 7.*12* (déc. 2011), p. 3353-3365 (cf. p. 86).

[Pi+15] Liya Pi et al., « A Disintegrin and Metalloprotease with Thrombospondin Type I Motif 7 », in : *The American Journal of Pathology* 185.*6* (juin 2015), p. 1552-1563, ISSN : 0002-9440, DOI : 10.1016/j.ajpath.2015.02.008, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450322/ (visité le 05/01/2022) (cf. p. 184, 218, 220).

[PMB04] M. Pagel, A. Meade et D. Barker, « Bayesian estimation of ancestral character states on phylogenies », in : *Syst Biol* 53.*5* (oct. 2004), p. 673-684 (cf. p. 56).

[PTM07] K. D. Pruitt, T. Tatusova et D. R. Maglott, « NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins », in : *Nucleic Acids Res* 35.*Database issue* (jan. 2007), p. D61-65 (cf. p. 29).

[Rau+21] C. Rauer et al., « Computational approaches to predict protein functional families and functional sites », in : *Curr Opin Struct Biol* 70 (oct. 2021), p. 108-122 (cf. p. 66).

[RC20] Vincent Ranwez et Nathalie N. Chantret, « Strengths and Limits of Multiple Sequence Alignment and Filtering Methods », in : *Phylogenetics in the Genomic Era*, sous la dir. de Celine Scornavacca, Frédéric Delsuc et Nicolas Galtier, No commercial publisher | Authors open access book, 2020, 2.2 :1-2.2 :36, URL : https://hal.archives-ouvertes.fr/hal-02535389 (cf. p. 47).

[Red+21] S. Redondo-García et al., « ADAMTS proteases and the tumor immune microenvironment : Lessons from substrates and pathologies », in : *Matrix Biol Plus* 9 (fév. 2021), p. 100054 (cf. p. 90).

[Rib+19] António J. M. Ribeiro et al., « Emerging concepts in pseudoenzyme classification, evolution, and signaling », en, in : *Science Signaling* 12.*594* (août 2019), eaat9797, ISSN : 1945-0877, 1937-9145, DOI : 10.1126/scisignal.aat9797, URL : https://stke.sciencemag.org/content/12/594/eaat9797 (visité le 28/10/2019) (cf. p. 39).

[Riv+10] S. Rivera et al., « Metzincin proteases and their inhibitors : foes or friends in nervous system physiology ? », in : *J Neurosci* 30.*46* (nov. 2010), p. 15337-15357 (cf. p. 79).

[RLB00] P. Rice, I. Longden et A. Bleasby, « EMBOSS : the European Molecular Biology Open Software Suite », in : *Trends Genet* 16.*6* (juin 2000), p. 276-277 (cf. p. 102).

[Ros+21] Keron W. J. Rose et al., « Regulation of ADAMTS Proteases », en, in : *Frontiers in Molecular Biosciences* 8 (juin 2021), p. 701959, ISSN : 2296-889X, DOI : 10.3389/fmolb.2021.701959, URL : https://www.frontiersin.org/articles/10.3389/fmolb.2021.701959/full (visité le 06/10/2021) (cf. p. 80, 82).

[RS08] R. H. Ree et S. A. Smith, « Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis », in : *Syst Biol* 57.*1* (fév. 2008), p. 4-14 (cf. p. 56).

[San+01] J. D. Sandy et al., « Versican V1 proteolysis in human aorta in vivo occurs at the Glu441-Ala442 bond, a site that is cleaved by recombinant ADAMTS-1 and ADAMTS-4 », eng, in : *The Journal of Biological Chemistry* 276.*16* (avr. 2001), p. 13372-13378, ISSN : 0021-9258, DOI : 10.1074/jbc.M009737200 (cf. p. 224, 228).

[San+19a] S. Santamaria et al., « Exosites in Hypervariable Loops of ADAMTS Spacer Domains control Substrate Recognition and Proteolysis », in : *Sci Rep* 9.*1* (juill. 2019), p. 10914 (cf. p. 228).

[San+19b] Salvatore Santamaria et al., « Exosites in Hypervariable Loops of ADAMTS Spacer Domains control Substrate Recognition and Proteolysis », in : *Scientific Reports* 9 (juill. 2019), p. 10914, ISSN : 2045-2322, DOI : 10.1038/s41598-019-47494-w, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6662762/ (visité le 01/03/2022) (cf. p. 226, 227).

[San+21] Salvatore Santamaria et al., « Post-translational regulation and proteolytic activity of the metalloproteinase ADAMTS8 », eng, in : *The Journal of Biological Chemistry* 297.*5* (nov. 2021), p. 101323, ISSN : 1083-351X, DOI : 10.1016/j.jbc.2021.101323 (cf. p. 92, 228).

[San20] Salvatore Santamaria, « ADAMTS-5 : A difficult teenager turning 20 », in : *International Journal of Experimental Pathology* 101.*1-2* (2020), p. 4-20, ISSN : 0959-9673, DOI : 10.1111/iep.12344, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7306899/ (visité le 05/01/2022) (cf. p. 184).

[Sau+17] H. Sauquet et al., « The ancestral flower of angiosperms and its early diversification », in : *Nat Commun* 8 (août 2017), p. 16047 (cf. p. 56).

[Say+19] Eric W. Sayers et al., « Database resources of the National Center for Biotechnology Information », eng, in : *Nucleic Acids Research* 47.*D1* (jan. 2019), p. D23-D28, ISSN : 1362-4962, DOI : 10.1093/nar/gky1069 (cf. p. 115).

[SBM16] Amarda Shehu, Daniel Barbará et Kevin Molloy, « A Survey of Computational Methods for Protein Function Prediction », in : *Big Data Analytics in Genomics*, sous la dir. de Ka-Chun Wong, Cham : Springer International Publishing, 2016, p. 225-298, ISBN : 978-3-319-41279-5, DOI : 10.1007/978-3-319-41279-5_7, URL : https://doi.org/10.1007/978-3-319-41279-5_7 (cf. p. 66, 76).

[Sch+18] Rahel Schnellmann et al., « A Selective Extracellular Matrix Proteomics Approach Identifies Fibronectin Proteolysis by A Disintegrin-like and Metalloprotease Domain with Thrombospondin Type 1 Motifs (ADAMTS16) and Its Impact on Spheroid Morphogenesis », in : *Molecular & Cellular Proteomics : MCP* 17.*7* (juill. 2018), p. 1410-1425, ISSN : 1535-9476, DOI : 10.1074/mcp.RA118.000676, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030725/ (visité le 05/01/2022) (cf. p. 184).

[Sch+20] Conrad L. Schoch et al., « NCBI Taxonomy : a comprehensive update on curation, resources and tools », eng, in : *Database : The Journal of Biological Databases and Curation* 2020 (jan. 2020), ISSN : 1758-0463, DOI : 10.1093/database/baaa062 (cf. p. 105, 109).

[Sch+86] T. D. Schneider et al., « Information content of binding sites on nucleotide sequences », in : *J Mol Biol* 188.*3* (avr. 1986), p. 415-431 (cf. p. 70).

[SG20] S. Santamaria et R. de Groot, « ADAMTS proteases in cardiovascular physiology and disease », in : *Open Biol* 10.*12* (déc. 2020), p. 200333 (cf. p. 89).

[SH21a] B. Satz-Jacobowitz et D. Hubmacher, « The quest for substrates and binding partners : A critical barrier for understanding the role of ADAMTS proteases in musculoskeletal development and disease », in : *Dev Dyn* 250.*1* (jan. 2021), p. 8-26 (cf. p. 91).

[SH21b] Brandon Satz-Jacobowitz et Dirk Hubmacher, « The quest for substrates and binding partners : A critical barrier for understanding the role of ADAMTS proteases in musculoskeletal development and disease », eng, in : *Developmental Dynamics : An Official Publication of the American Association of Anatomists* 250.*1* (jan. 2021), p. 8-26, ISSN : 1097-0177, DOI : 10.1002/dvdy.248 (cf. p. 184).

[Sha+03] P. Shannon et al., « Cytoscape : a software environment for integrated models of biomolecular interaction networks », in : *Genome Res* 13.*11* (nov. 2003), p. 2498-2504 (cf. p. 186).

[Sie+11] F. Sievers et al., « Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega », in : *Mol Syst Biol* 7 (oct. 2011), p. 539 (cf. p. 140, 141).

[Sig+02] Christian J. A. Sigrist et al., « PROSITE : a documented database using patterns and profiles as motif descriptors », eng, in : *Briefings in Bioinformatics* 3.*3* (sept. 2002), p. 265-274, ISSN : 1467-5463, DOI : 10.1093/bib/3.3.265 (cf. p. 69, 152).

[SM87] David L Swofford et Wayne P Maddison, « Reconstructing ancestral character states under Wagner parsimony », in : *Mathematical biosciences* 87.*2* (1987), p. 199-229 (cf. p. 56).

[Sma+21] F. Z. Smaili et al., « QAUST : Protein Function Prediction Using Structure Similarity, Protein Interaction, and Functional Motifs », in : *Genomics Proteomics Bioinformatics* 19.*6* (déc. 2021), p. 998-1011 (cf. p. 66).

[SOG19] B. Smithers, M. Oates et J. Gough, « 'Why genes in pieces ?'-revisited », in : *Nucleic Acids Res* 47.*10* (juin 2019), p. 4970-4973 (cf. p. 58, 233).

[Som+03] Robert P. T. Somerville et al., « Characterization of ADAMTS-9 and ADAMTS-20 as a distinct ADAMTS subfamily related to Caenorhabditis elegans GON-1 », eng, in : *The Journal of Biological Chemistry* 278.*11* (mars 2003), p. 9503-9513, ISSN : 0021-9258, DOI : 10.1074/jbc.M211009200 (cf. p. 184, 224, 228).

[Spe+21] M. A. Spence et al., « Ancestral sequence reconstruction for protein engineers », in : *Curr Opin Struct Biol* 69 (août 2021), p. 131-141 (cf. p. 56).

[SR09] R. A. Studer et M. Robinson-Rechavi, « How confident can we be that orthologs are similar, but paralogs differ ? », in : *Trends Genet* 25.*5* (mai 2009), p. 210-216 (cf. p. 45).

[Sta+05] Heather Stanton et al., « ADAMTS5 is the major aggrecanase in mouse cartilage in vivo and in vitro », eng, in : *Nature* 434.*7033* (mars 2005), p. 648-652, ISSN : 1476-4687, DOI : 10.1038/nature03417 (cf. p. 92, 224, 228).

[Sta+20] Moses Stamboulian et al., « The ortholog conjecture revisited : the value of orthologs and paralogs in function prediction », en, in : *Bioinformatics* 36.*Supplement_1* (juill. 2020), Publisher : Oxford Academic, p. i219-i226, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/btaa468, URL : https://academic.oup.com/bioinformatics/article/36/Supplement_1/i219/5870499 (visité le 14/08/2020) (cf. p. 45, 191).

[Sta14] Alexandros Stamatakis, « RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies », in : *Bioinformatics* 30.*9* (mai 2014), p. 1312-1313, ISSN : 1367-4803, DOI : 10.1093/bioinformatics/btu033, URL : https://doi.org/10.1093/bioinformatics/btu033 (visité le 10/01/2022) (cf. p. 128).

[Sto+12] M. Stolzer et al., « Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees », in : *Bioinformatics* 28.*18* (sept. 2012), p. i409-i415 (cf. p. 60).

[Sto+15] M. Stolzer et al., « Event inference in multidomain families with phylogenetic reconciliation », in : *BMC Bioinformatics* 16 Suppl 14 (2015), S8 (cf. p. 58, 60, 159).

[SW12] K. Scheffzek et S. Welti, « Pleckstrin homology (PH) like domains - versatile modules in protein-protein interaction platforms », in : *FEBS Lett* 586.*17* (août 2012), p. 2662-2673 (cf. p. 37).

[Szk+19] D. Szklarczyk et al., « STRING v11 : protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets », in : *Nucleic Acids Res* 47.*D1* (jan. 2019), p. D607-D613 (cf. p. 237).

[Thé+21] Nathalie THÉRET et al., « ADAM and ADAMTS Proteins, New Players in the Regulation of Hepatocellular Carcinoma Microenvironment », eng, in : *Cancers* 13.*7* (mars 2021), ISSN : 2072-6694, DOI : 10.3390/cancers13071563 (cf. p. 79, 81, 99).

[THL11] A. TOFIGH, M. HALLETT et J. LAGERGREN, « Simultaneous identification of duplications and lateral gene transfers », in : *IEEE/ACM Trans Comput Biol Bioinform* 8.*2* (2011), p. 517-535 (cf. p. 60).

[Tho+11] Julie D. THOMPSON et al., « A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods : Current Challenges and Future Perspectives », en, in : *PLOS ONE* 6.*3* (mars 2011), e18093, ISSN : 1932-6203, DOI : 10.1371/journal.pone.0018093, URL : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018093 (visité le 08/11/2019) (cf. p. 47).

[TKL97] R. L. TATUSOV, E. V. KOONIN et D. J. LIPMAN, « A genomic perspective on protein families », in : *Science* 278.*5338* (oct. 1997), p. 631-637 (cf. p. 45, 190).

[TMK19] A. D. THEOCHARIS, D. MANOU et N. K. KARAMANOS, « The extracellular matrix as a multitasking player in disease », in : *FEBS J* 286.*15* (août 2019), p. 2830-2869 (cf. p. 89).

[Tor+02] Micky D. TORTORELLA et al., « Characterization of human aggrecanase 2 (ADAM-TS5) : substrate specificity studies and comparison with aggrecanase 1 (ADAM-TS4) », eng, in : *Matrix Biology : Journal of the International Society for Matrix Biology* 21.*6* (oct. 2002), p. 499-511, ISSN : 0945-053X, DOI : 10.1016/s0945-053x(02)00069-0 (cf. p. 224).

[Tor+22] C. TORET et al., « The cellular slime mold Fonticula alba forms a dynamic, multicellular collective while feeding on bacteria », in : *Curr Biol* 32.*9* (mai 2022), p. 1961-1973 (cf. p. 113).

[TOT02] Annabel E TODD, Christine A ORENGO et Janet M THORNTON, « Sequence and Structural Differences between Enzyme and Nonenzyme Homologs », en, in : *Structure* 10.*10* (oct. 2002), p. 1435-1451, ISSN : 0969-2126, DOI : 10.1016/S0969-2126(02)00861-4, URL : http://www.sciencedirect.com/science/article/pii/S0969212602008614 (visité le 28/10/2019) (cf. p. 39).

[Tsu+10] K. Tsutsui et al., « ADAMTSL-6 is a novel extracellular matrix protein that binds to fibrillin-1 and promotes fibrillin-1 fibril formation », in : *J Biol Chem* 285.7 (fév. 2010), p. 4870-4882 (cf. p. 80).

[Van+16] Renaud Vanhoutreve et al., « LEON-BIS : multiple alignment evaluation of sequence neighbours using a Bayesian inference system », in : *BMC Bioinformatics* 17 (juill. 2016), ISSN : 1471-2105, DOI : 10.1186/s12859-016-1146-y, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4936259/ (visité le 08/11/2019) (cf. p. 234).

[Var+22] M. Varadi et al., « AlphaFold Protein Structure Database : massively expanding the structural coverage of protein-sequence space with high-accuracy models », in : *Nucleic Acids Res* 50.*D1* (jan. 2022), p. D439-D444 (cf. p. 34, 85-88).

[Wan+11] L. S. Wang et al., « The impact of multiple protein sequence alignment on phylogenetic estimation », in : *IEEE/ACM Trans Comput Biol Bioinform* 8.*4* (2011), p. 1108-1119 (cf. p. 140).

[Wan+19a] Lauren W. Wang et al., « A disintegrin-like and metalloproteinase domain with thrombospondin type 1 motif 9 (ADAMTS9) regulates fibronectin fibrillogenesis and turnover », in : *The Journal of Biological Chemistry* 294.*25* (juin 2019), p. 9924-9936, ISSN : 0021-9258, DOI : 10.1074/jbc.RA118.006479, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6597835/ (visité le 05/01/2022) (cf. p. 184).

[Wan+19b] Lauren W. Wang et al., « Adamts10 inactivation in mice leads to persistence of ocular microfibrils subsequent to reduced fibrillin-2 cleavage », in : *Matrix biology : journal of the International Society for Matrix Biology* 77 (avr. 2019), p. 117-128, ISSN : 0945-053X, DOI : 10.1016/j.matbio.2018.09.004, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8209899/ (visité le 05/01/2022) (cf. p. 89, 184).

[Wei+18] Jian-lu Wei et al., « ADAMTS-12 protects against inflammatory arthritis through interacting with and inactivating proinflammatory CTGF », in : *Arthritis & rheumatology (Hoboken, N.J.)* 70.*11* (nov. 2018), p. 1745-1756, ISSN : 2326-5191, DOI : 10.1002/art.40552, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6203634/ (visité le 10/01/2022) (cf. p. 218, 220).

[Wer+14] G. D. Werner et al., « A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms », in : *Nat Commun* 5 (juin 2014), p. 4087 (cf. p. 56).

[Wil+21] T. A. Williams et al., « Inferring the Deep Past from Molecular Data », in : *Genome Biol Evol* 13.*5* (mai 2021) (cf. p. 51).

[WM17] S. Whelan et D. A. Morrison, « Inferring Trees », in : *Methods Mol Biol* 1525 (2017), p. 349-377 (cf. p. 52-54).

[WRK12] Y. C. Wu, M. D. Rasmussen et M. Kellis, « Evolution at the subgene level : domain rearrangements in the Drosophila phylogeny », in : *Mol Biol Evol* 29.*2* (fév. 2012), p. 689-705 (cf. p. 58, 157, 233).

[Wu+13] Yi-Chieh Wu et al., « TreeFix : Statistically Informed Gene Tree Error Correction Using Species Trees », in : *Systematic Biology* 62.*1* (jan. 2013), p. 110-120, ISSN : 1063-5157, DOI : 10.1093/sysbio/sys076, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3526801/ (visité le 23/10/2019) (cf. p. 128, 161).

[Yam+14] Kazuhiro Yamamoto et al., « Low density lipoprotein receptor-related protein 1 (LRP1)-mediated endocytic clearance of a disintegrin and metalloproteinase with thrombospondin motifs-4 (ADAMTS-4) : functional differences of non-catalytic domains of ADAMTS-4 and ADAMTS-5 in LRP1 binding », eng, in : *The Journal of Biological Chemistry* 289.*10* (mars 2014), p. 6462-6474, ISSN : 1083-351X, DOI : 10.1074/jbc.M113.545376 (cf. p. 184, 224).

[Yan+15] J. Yang et al., « The I-TASSER Suite : protein structure and function prediction », in : *Nat Methods* 12.*1* (jan. 2015), p. 7-8 (cf. p. 34).

[Yan97] Z. Yang, « PAML : a program package for phylogenetic analysis by maximum likelihood », in : *Comput Appl Biosci* 13.*5* (oct. 1997), p. 555-556 (cf. p. 55).

[Yon+20] George G. Vega Yon et al., « On the automatic annotation of gene functions using observational data and phylogenetic trees », en, in : *bioRxiv* (mai 2020), Publisher : Cold Spring Harbor Laboratory Section : New Results, p. 2020.05.14.095687, DOI : 10.1101/2020.05.14.095687, URL : https://www.biorxiv.org/content/10.1101/2020.05.14.095687v1 (visité le 15/09/2020) (cf. p. 56, 76, 201).

[Yu+19] Lijia Yu et al., « Grammar of protein domain architectures », in : *Proceedings of the National Academy of Sciences of the United States of America* 116.*9* (fév. 2019), p. 3636-3645, ISSN : 0027-8424, DOI : 10.1073/pnas.1814684116, URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6397568/ (visité le 27/11/2019) (cf. p. 36).

[Zea+21] D. J. Zea et al., « Assessing conservation of alternative splicing with evolutionary splicing graphs », in : *Genome Res* 31.*8* (août 2021), p. 1462-1473 (cf. p. 236).

UNIVERSITÉ DE RENNES 1

Titre : Caractérisation en modules fonctionnels des protéines ADAMTS-TSL, par approches de phylogénies

Mot clés : ADAMTS-TSL, phylogénomique, région conservée, réconciliation phylogénétique, protéines multidomaines, annotation fonctionnelle, interaction protéine-protéine

Résumé : Les protéines multidomaines ADAMTS-TSL humaines sont impliquées dans de nombreuses pathologies. Codées par 26 gènes paralogues, leur combinatoire en domaines ne suffit pas à caractériser leurs différences fonctionnelles. Nous proposons dans cette thèse une nouvelle approche d'identification des régions fonctionnelles des séquences. Pour cela nous utilisons des séquences de 9 espèces eucaryotes afin d'identifier des modules de séquences conservées propres à certains sous-groupes de séquences homologues. L'analyse évolutive des modules identifiés est obtenue en effectuant une reconstruction phylogénétique conjointe des gènes, des espèces et des modules. Par ailleurs, pour valider l'intérêt fonctionnel des modules identifiés, nous associons des phénotypes (PPI) à cette histoire évolutive. Ce qui a abouti à identifier des acquisitions concomitantes de « modules/phénotypes », prédisant la fonctionnalité de ces modules. Appliquer cette approche aux protéines ADAMTS-TSL humaines nous a permis d'identifier de nouvelles régions fonctionnelles, plus fines, non contiguës et à même d'en décrire les spécificités.

Title: Characterization in functional modules of ADAMTS-TSL proteins, by phylogeny approaches

Keywords: ADAMTS-TSL, phylogenomics, conserved region, phylogenetic reconciliation, multidomain proteins, functional annotation, protein-protein interaction

Abstract: The human ADAMTS-TSL multidomain proteins are involved in numerous pathologies. Encoded by 26 paralogous genes, their domain combination is not sufficient to characterize their functional differences. We propose in this thesis a new approach to identify functional regions of the sequences. For this purpose, we use sequences from 9 eukaryotic species to identify conserved sequence modules specific to certain subgroups of homologous sequences. The evolutionary analysis of the identified modules is obtained by performing a joint phylogenetic reconstruction of genes, species and modules. Furthermore, to validate the functional interest of the identified modules, we associate phenotypes (PPI) to this evolutionary history. This has led to the identification of concomitant acquisitions of "modules/phenotypes", predicting the functionality of these modules. Applying this approach to human ADAMTS-TSL proteins has allowed us to identify new, finer, non-contiguous functional regions that can describe their specificities.