Recherche de motifs fonctionnels dans une famille de protéines multi-domaines : les Fibulines

Elisa Chenel

▶ **To cite this version:**

Elisa Chenel. Recherche de motifs fonctionnels dans une famille de protéines multi-domaines : les Fibulines. Bio-informatique [q-bio.QM]. 2024. hal-04839500

HAL Id: hal-04839500
https://inria.hal.science/hal-04839500v1

Submitted on 17 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

UNIVERSITE DE RENNES
MASTER 2 BIO-INFORMATIQUE
Parcours informatique pour la biologie et la santé
2023-2024

Recherche de motifs fonctionnels dans une famille de protéines multi-domaines : les Fibulines

---

Elisa Chenel

Sous la direction de Nathalie Théret (DR Inserm) et Samuel Blanquart (CR Inria)

Equipe Dyliss IRISA - Campus de Beaulieu 263 avenue du Général Leclerc 35000 Rennes

DYMEC2

Date d'écriture : 2024-06-26
Soutenance : 2024-07-03

Université de Rennes

UFR Sciences de la vie & de l'environnement

ENGAGEMENT DE NON PLAGIAT

Je, soussigné (e) Elisa CHENEL

Etudiant (e) en Master 2 Bio-informatique : parcours informatique pour la biologie et la santé

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

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

Signature

INSERM U1242 OSS Equipe PROSAC

Centre Eugène Marquis Avenue de la Bataille Flandres Dunkerque
35042 Rennes

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr

TÉL. 33 (0)2 23 23 61 14

Remerciements

Je tiens tout d'abord à remercier mes maitres de stage, Nathalie THERET et Samuel BLANQUART, ainsi qu'Oliver DENNLER, François COSTE et Catherine BELLEANNEE, pour leurs accompagnements tout au long de mon stage, ainsi que pour leurs conseils et explications qui m'ont permis de mener à bien mon stage. Les remarques et suggestions faites lors de l'écriture de mon rapport m'ont également beaucoup apportées.
Je souhaite remercier l'IRISA, l'équipe DYLISS, l'IREST et l'équipe DYMEC2 de m'avoir accueillie pour mon stage de 6 mois.
Je veux aussi remercier mes voisins de bureau : Juliette, Lune, Melody, Hugo et Alix, pour leurs aides, mais aussi pour les pauses partagées ensemble.
Et enfin, merci à ma famille qui m'a soutenue tout au long de mon stage.

Sommaire

1 Introduction

1.1 Contexte biologique

1.1.1 Les maladies chroniques hépatiques et la matrice extracellulaire

Ce projet s'inscrit dans le cadre des recherches effectuées par l'équipe dirigée par Nathalie Théret et qui portent sur la dynamique du microenvironnement au cours de la progression des maladies chroniques du foie conduisant au cancer. Ces maladies sont principalement liées aux hépatites virales, à la consommation d'alcool et aux perturbations métaboliques (diabète, obésité, etc.). Elles se caractérisent par le développement d'une fibrose qui consiste en une accumulation excessive de matrice extracellulaire qui va altérer les fonctions hépatiques. La matrice extracellulaire est une structure tridimensionnelle dynamique qui entoure toutes les cellules. Elle constitue un réseau moléculaire complexe dans lequel interagit une grande variété de protéines (Figure 1).

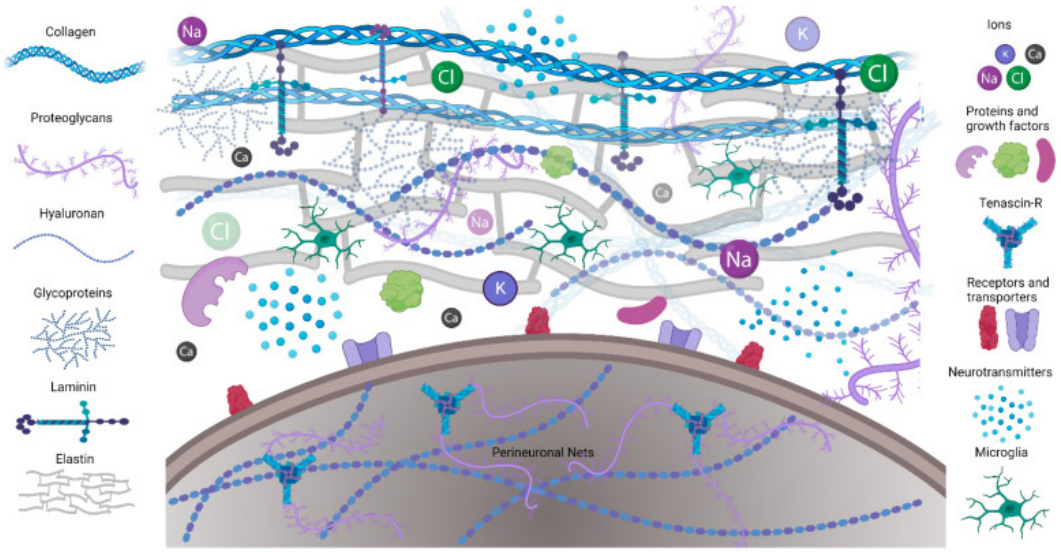


Figure 1. **Schéma illustrant la matrice extracellulaire et ses composants**. Les différents composants de la matrice extracellulaire permettent un soutien structurel et permettent de créer un micro-environnement favorable aux interactions entre les cellules. (d'après [1])

Afin de caractériser les modifications de la matrice extracellulaire dans les maladies chroniques hépatiques, l'équipe DYMEC2 de l'IRSET a récemment réalisé une analyse protéomique de tissus fibreux chez des patients atteints de carcinome hépatocellulaire, un cancer du foie. Trois membres de la famille des fibulines ont été identifiés et associés à la gravité de la fibrose, cependant les fonctions de ces fibulines dans la pathologie sont très peu connues. Dans ce contexte, l'équipe a souhaité développer une approche globale de recherche de motifs fonctionnels au sein de cette famille de protéines.

1.1.2 Les fibulines, une famille de protéines de la matrice extracellulaire

Les fibulines sont une famille de protéines de la matrice extracellulaire composée de huit protéines humaines [2]. La structure de ces protéines est organisée en trois régions (Figure 2). La première région correspond à la partie N-terminale des protéines dont la longueur varie suivant les fibulines et présente différents domaines qui peuvent être spécifiques à certaines fibulines. La deuxième région correspond à la partie centrale des protéines et se caractérise par un nombre variable de motifs de type EGF-like (Pfam, IPR000742) (EGF : *epidermal growth factor*) et de motifs de type EGF-like calcium-binding (Pfam,IPR001881). Enfin, la troisième région contient le domaine de type fibuline (”Fibulin Type Module” figure 2) (Pfam, IPR017048) en position C-terminale avec une longueur variant de 120 à 140 acides aminés.

Ces protéines sont classées en trois sous-groupes. Les fibulines, dites ”longues”, correspondent aux fibulines-1 et -2. Outre la présence de motifs de type EGF-like calcium-binding, ces fibulines se caractérisent par la présence de motifs de type Anaphylatoxin-like (Pfam, IPR000020). Les fibulines, dites ”courtes”, correspondent aux fibulines-3-4-5 et -7, cette dernière se différencie par la présence d'un motif de type Shushi (Pfam, IPR053298). Enfin, le dernier sous-groupe est composé des hémicentines-1 et -2, correspondant respectivement aux fibulines-6 et -8. Elles se caractérisent par un motif de type von Willebrand A (Pfam, IPR013608) à l'extrémité N-terminale et une longue série (>40) de motifs de type Immunoglobulines (Pfam, IPR008424).
Le rôle des fibulines est associé à leur capacité à interagir avec d'autres composants matriciels. Elles stabilisent ainsi l'organisation supramoléculaire des fibres élastiques et des fibres de collagène et contribuent à l'intégrité structurelle de la membrane basale. Ce positionnement au sein de la matrice en font des modulateurs du remodelage matriciel qui affecte le phénotype cellulaire (migration, prolifération). Cependant, il n'existe que très peu d'information quant aux séquences impliquées dans ces fonctions, ce qui limite le développement d'un ciblage thérapeutique de ces protéines.

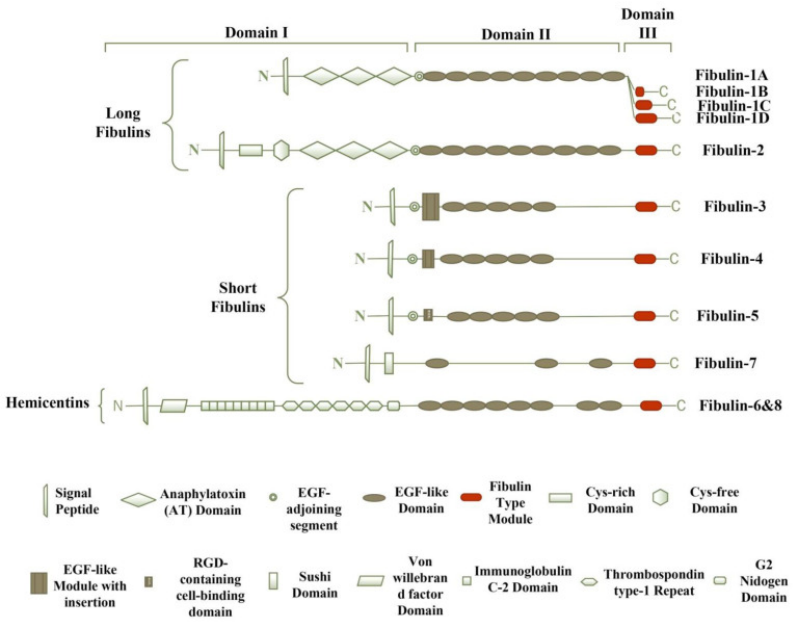


Figure 2. **Schéma de la structure moléculaire des protéines de la famille des fibuline humaines.** La famille des fibulines est composée de huit membres, chacune est constituées de trois domaines (I, II, III). L'ensemble des fibulines présente des domaines EGF-like, ainsi qu'un ”fibuline-like module”. (d'après [2])

1.2 Contexte informatique : annotation fonctionnelle des protéines et prédiction de motifs

L'annotation fonctionnelle des protéines et plus particulièrement la caractérisation des séquences impliquées dans une fonction est indispensable pour le développement d'approches de ciblage thérapeutique. De très nombreuses méthodes de prédiction de fonction des protéines ont été développées et ont fait l'objet de revues récentes [3, 4, 5]. Ces méthodes font appel à différents types de données, notamment les séquences d'acides aminés et les structures tridimensionnelles. Plus récemment, les données issues des technologies à haut débit comme les profils d'expression (protéines ou ARN) et les interactions moléculaires fournissent de nouvelles sources de données et permettent d'améliorer les méthodes de prédictions. Nous ne détaillerons dans ce manuscrit que les principales approches basées sur la structure et la séquence. Ces méthodes se basent sur l'hypothèse que s'il existe une ressemblance entre les séquences ou les structures des protéines, alors celles-ci pourraient partager une même fonction. L'outil utilisé pour notre étude a fait appel aux méthodes d'homologie de séquence, de phylogénétique et de réconciliation.

1.2.1 Approches basées sur la structure

Certaines caractéristiques "fonctionnelles" sont inscrites dans la structure tridimensionnelle des protéines, par exemple les structures en hélice alpha composées de 20 à 30 acides aminés hydrophobes qui caractérisent les protéines transmembranaires, les sites actifs des enzymes qui forment une région d'interaction pour un substrat ou encore les sites spécifiques d'interactions protéine-protéine comme dans les complexes ligand-récepteur. Parmi les méthodes qui ont été développées, certaines se basent sur des algorithmes d'alignement structural, qu'il soit global ou local. Ces méthodes utilisent les caractéristiques de repliement de protéines dont la fonction est connue pour annoter de nouvelles protéines non annotées par similarité de structure [6].
L'article [7] propose une méthode permettant la prédiction de sites de liaison pour des petites molécules. Cette méthode, basée sur la structure des protéines, intègre aussi des estimations de la conservation évolutive des séquences afin d'identifier les cavités à la surface des protéines.
Ces méthodes ne sont utilisables que pour des structures statiques. L'utilisation de la dynamique structurelle peut permettre d'améliorer la prédiction des fonctions. Pour prendre en compte cette information, une simulation de la dynamique moléculaire via l'utilisation d'algorithmes de prédiction basés sur la structure permet d'identifier les sites de liaison et donc la prédiction de fonctions [8].
Si l'utilisation des structures est intéressante pour l'annotation fonctionnelle des protéines, ces méthodes sont restées longtemps limitées en raison du manque de données structurelles à haute résolution. Cependant, l'arrivée récente des méthodes de deeplearning a révolutionné les méthodes de prédiction des structures [9] avec tout particulièrement l'arrivée du système AlphaFold, développé par Google DeepMind [10].

1.2.2 Approches basées sur les séquences

Homologie de séquences
La méthode par homologie de séquence se base sur l'hypothèse que si deux séquences sont similaires, elles dérivent d'un ancêtre commun et possèdent donc des fonctions similaires. L'un des moyens de modéliser cette conservation de séquence est de réaliser un alignement de séquences multiples (MSA). Il existe trois types de méthodes pour la modélisation d'un MSA. La première se base sur des motifs correspondant à des expressions régulières. Cet alignement permet de mettre en avant des motifs, indiquant les variations possibles pour les différentes positions ainsi

que les positions les séparant. Les motifs Prosite [11] sont un exemple de ces motifs.

La deuxième méthode de modélisation est réalisée par Position-Specific Scoring System (PSSM) [12]. Elle correspond à une matrice composée de poids où un poids est attribué à chaque acide aminé. Ce poids correspond au log du rapport de vraisemblance entre la fréquence observée et celle attendue [3]. Il est possible de représenter plusieurs régions conservées au sein d'une même famille via un enchaînement ordonné de PSSM modélisant ainsi une signature appelée fingerprints [13]. Cette approche ne permet pas de prendre en compte les insertions et les délétions.

La dernière méthode utilise ce qu'on appelle un profil HMM (pHMM : *profile Hidden Markov models*). Cette approche permet une représentation probabiliste d'un MSA, par l'utilisation d'un modèle statistique permettant de modéliser les conservations ainsi que les insertions et les délétions, contrairement à la méthode PSSM. Les domaines présents dans la base de données Pfam [14] ont été identifiés par cette approche.

Au cours de notre étude, nous avons utilisé la suite Protomata développée par François Coste [15], [16] qui permet l'apprentissage de la signature spécifique d'une famille de protéines homologues à partir de séquences non-alignées par alignement multiple partiel et local et par modélisation d'automate. Un des programmes présents dans la suite Protomata est appelé Paloma. Il permet la construction des alignements multiples locaux partiels (PLMA : *Partial Local Multiple Alignment*) (Figure 3) permettant d'identifier les régions localement conservées. Ces régions sont alors appelées blocs de conservation. L'identification de ces blocs de conservation permet de caractériser une famille de protéines tout en considérant l'hétérogénéité des séquences entre elles. Un bloc est alors constitué d'un sous-groupe de séquences (au minimum deux). Une analyse Paloma réalise, à partir de l'ensemble des paires d'alignement local, une intégration itérative par alignement transitif des positions est faite avec les paires d'alignement locales. Si les blocs d'alignement issus de Paloma sont assez longs, ils seront considérés par la suite comme des modules de conservation.

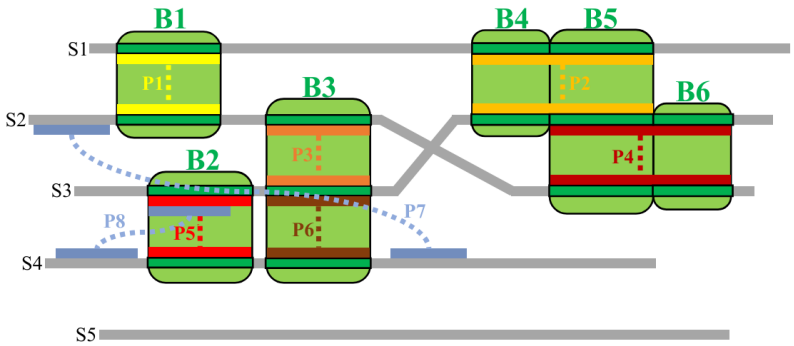


Figure 3. **Identification des modules par alignement multiple local partiel (PLMA).** Schématisation d'un PLMA de cinq séquences S1...S5, composé de six blocs d'alignement B1,..., B6. Chaque bloc est réalisé à partir d'un alignement local des séquences pour certaines positions des séquences. Cet alignement local peut être visualisé par exemple avec le bloc B2 où deux segments nommés P5 présent chez les séquences S3 et S4 s'alignent sur ces séquences. Le segment P8 représente aussi un alignement entre les séquences S3 et S4, sa taille étant inférieure à celle du segment P5 il n'a pas été sélectionné par Paloma. Ces blocs ne sont pas spécifiques à des paires de séquences comme le bloc B3 qui réalise des alignements locaux avec des segments des séquences 2, 3 et 4. Le bloc B4, B5 et B6 sont un exemple des découpages qui peuvent être réalisés lors de l'identification de bloc. Pour deux paires d'alignement (P2 et P4), trois blocs sont identifiés : le bloc B4 correspondant à un alignement entre S1 et S3, le bloc B5 pour les positions contiguës entre les séquences S1, S2 et S3 et enfin un bloc B6 pour celles entre S2 et S3(d'après [17, 18])

Phylogénomique

L'approche basée sur la phylogénomique pour l'inférence de fonctions biologiques a été initialement décrite par Jonathan Eisen [19]. Cette approche se base sur plusieurs étapes impliquant la sélection de séquences partageant une origine commune, ces séquences sont dites homologues, l'alignement de séquences multiples (MSA) et la construction d'un arbre phylogénétique, la superposition d'annotations sur les topologies de l'arbre, la distinction entre orthologues (séquences homologues issues d'une spéciation) et paralogues (séquences homologues issues d'une duplication), et enfin l'inférence de la fonction d'une protéine sur la base des orthologues identifiés par ce processus et des annotations récupérées. Il existe trois méthodes de reconstruction phylogénétique qui peuvent être utilisées pour l'inférence phylogénétique.
Ces méthodes se basent sur l'hypothèse que deux séquences proches dans un arbre phylogénétique sont susceptibles de partager une même fonction.

La première est appelée méthode de distance. Pour celle-ci, une matrice des distances est construite à partir de la matrice des caractères. La matrice de distances représente alors les distances évolutives entre toutes les paires de protéines. Ensuite, l'arbre phylogénétique est déduit de cette matrice à l'aide d'algorithmes.

La deuxième méthode est la méthode de minimum parsimony, elle se base sur la sélection de l'arbre nécessitant le plus petit nombre de changements afin d'expliquer les données observées.

La dernière est la méthode de vraisemblance se basant sur une fonction permettant de calculer la probabilité qu'un arbre ait pu produire les données observées [20]. Ces trois approches produisent la phylogénie qui peut être utilisée par des méthodes d'annotation phylogénomique.

1.3 Phylogénie moléculaire

La phylogénie moléculaire tend à reconstruire l'évolution de ces séquences. L'objectif premier est alors de déterminer l'arbre phylogénétique qui permet de décrire les relations de parenté évolutive des séquences homologues à partir d'ADN, d'ARN ou de protéines.

1.3.1 Méthodes de réconciliation

La réconciliation est une approche qui permet de connecter l'histoire évolutive de deux ou plus entités biologiques co-évoluant. Un arbre phylogénétique représentant l'évolution d'une entité peut alors être "dessiné" au sein d'un autre arbre phylogénétique représentant l'histoire évolutive d'une autre entité. Cette approche permet de mettre en évidence l'interdépendance des deux entités et les différents événements évolutifs qui ont marqué leur histoire commune. Ces méthodes de réconciliation sont adaptées à la problématique des protéines multidomaines. L'histoire des gènes est parfois indépendante de celle des espèces, en raison des événements de duplication et de perte de gènes. De façon similaire, l'évolution de ces protéines peut se faire par brassage de domaines, ce qui inclut différents événements tels que les insertions, les transferts, les duplications, la fusion et la perte de domaines. Ce brassage de domaine peut entraîner des différences de fonctions des protéines. Pour comprendre l'évolution de ces protéines multidomaines, il y a donc besoin de comprendre l'évolution de l'architecture des domaines. Stolzer [21] a proposé une méthode de réconciliation Domaine-Gène chez les familles de protéines multidomaines. Les entités ici réconciliées sont un arbre des gènes et un arbre des domaines. La réconciliation aura pour but d'associer les nœuds de chaque arbre entre eux. Cette approche, bien qu'innovante, ne prend pas en compte l'histoire des espèces. Li et Bansal [22] ont récemment proposé de réaliser la réconciliation entre les trois niveaux : "Domaines, Gènes et Espèces" (DGS) avec le développement de l'outil SEADOG. Pour ce faire, deux types de réconciliation sont réalisés, une réconciliation Gène-Espèce et une réconciliation Domaine-Gène. Une extension de cette méthode, SEADOG-MD permet aujourd'hui de prendre en compte la présence de plusieurs domaines, caractéristiques propres aux protéines multidomaines [23] et permet de considérer les réconciliations Gène-Espèce, mais aussi l'ensemble des réconciliations Domaine-Gène. Cette méthode se base sur un alignement de séquences multiples, un arbre phylogénétique est alors réalisé pour chaque module.

1.3.2 Approche phylogénomique

C'est en se basant sur cette réconciliation à trois niveaux qu'Olivier Dennler a développé l'outil PhyloCharMod [18] (Phylogenetic Characterization of Modules). Cet outil permet de caractériser des motifs fonctionnels par la détection de modules conservés en utilisant l'inférence phylogénétique de l'histoire évolutive des espèces, des gènes, des modules et des fonctions. L'outil PhyloCharMod se décompose en différentes étapes (Figure 4). La première étape est le calcul de l'arbre des gènes à partir des séquences données en entrée de l'outil. Cette étape comprend le calcul d'un alignement de séquences multiples (MSA) avec l'outil Muscle [24] suivi d'une inférence de la phylogénie avec l'outil PhyML [25] et d'une correction de l'arbre avec l'outil TreeFix [26]. La deuxième étape a pour but d'identifier des modules conservés par l'utilisation de l'outil Paloma [16]. Comme décrit ci-dessus, cet outil permet de déterminer des ensembles de segments similaires d'au moins deux séquences protéiques alignées dans un PLMA. Ces ensembles sont alors nommés modulent. Ces derniers se différencient des domaines qui eux correspondent à des régions protéiques auxquelles sont associées des annotations (fonction). La troisième étape permet d'inférer la composition en modules des gènes ancestraux. Cette étape

se base sur la méthode de réconciliation module-gène-espèce réalisée avec l'outil SEADOG-MD [23]. L'étape quatre consiste en l'annotation des phénotypes connus des protéines. Dans notre étude, les phénotypes considérés sont les interactions protéine-protéine (PPI), l'hypothèse étant que l'interaction entre deux protéines signe une fonctionnalité (par exemple : enzyme et substrat, ligand et récepteur). L'étape cinq réalise une reconstruction des scénarios ancestraux de l'évolution des phénotypes le long de l'arbre des gènes en utilisant l'outil PastML [27]. La dernière étape réalise un regroupement des données des modules et des phénotypes. Cette étape permet d'obtenir pour chaque gène ancestral sa composition en module et ses traits phénotypiques (les PPI dans notre cas).

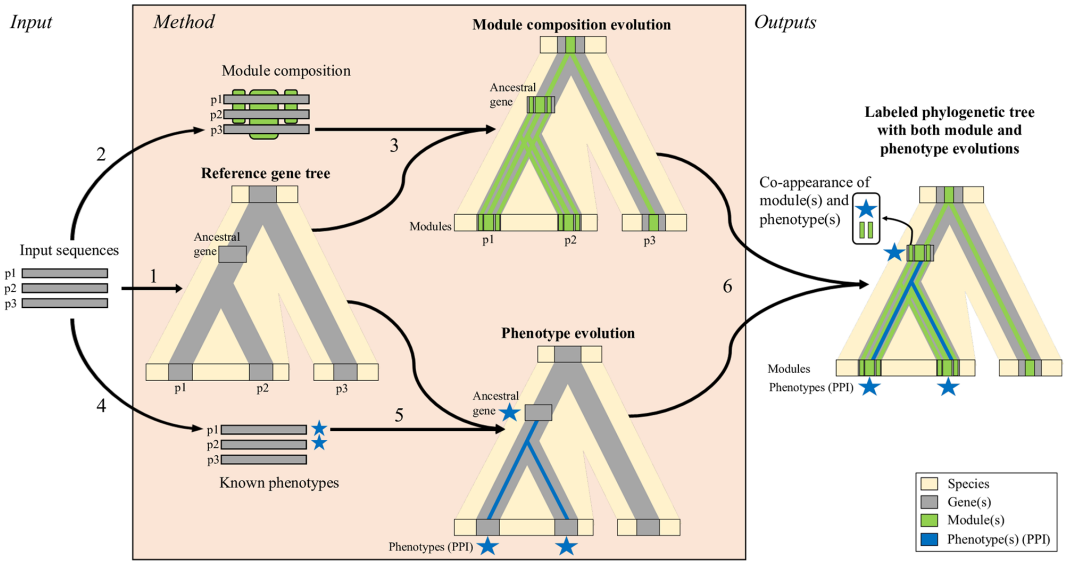


Figure 4. **Schéma de l'outil PhyloCharMod réalisant une inférence de l'évolution conjointe des modules et des phénotypes.** Cet outil réalise différentes étapes 1) Inférence de l'arbre de gène (*Muscle*,*PhyML*, *TreeFix*), 2) identification de modules conservés (*Paloma*), 3) inférence de la composition en modules pour les gènes ancestraux (*SEADOG-MD*), 4) Annotation des protéines avec les traits phénotypiques connus, 5) Reconstruction des scénarios ancestraux des phénotypes (*PastML*), 6) regroupement des données. Les résultats issus de l'outil permettent une estimation des événements de co-apparition. Ces derniers correspondent à un gain simultané d'un module et d'un phénotype chez un gène ancestral. (d'après [18])

1.4 Objectifs du projet

L'objectif principal de mon stage était la recherche de motifs fonctionnels chez la famille de protéines multidomaines : les fibulines. Pour réaliser cette étude, j'ai utilisé l'outil PhyloCHarMod qui a été développé pour identifier des modules conservés et utilise l'évolution associée aux modules conservés afin de les associer à une fonction.
Outre cette étude des fibulines, un des objectifs du stage était d'une part de savoir si l'outil PhyloCHarMod conçu initialement pour les protéines multidomaines, ADAMTS-TSL, était généralisable à d'autres familles de protéines multidomaines de la matrice extra cellulaire et d'autre part de rendre l'outil utilisable sur la plateforme Genouest.

2 Matériel et méthode

2.1 Un pipeline d'annotation phylogénomique

PhyloCharMod est un outil développé pour caractériser des motifs fonctionnels au sein de protéines multidomaines et de les associer à des phénotypes. L'outil se base sur une méthode par homologie et phylogénomique [17, 18].

2.1.1 Utilisation de l'outil

PhyloCharMod prend obligatoirement un fichier fasta contenant les séquences et un fichier d'annotation. La commande suivante est alors utilisée pour lancer le pipeline :
`python3 /phylocharmod/phylocharmod.py multi_fasta_file leaf_functions_csv`
L'outil propose de donner en entrée un arbre des gènes, un arbre des espèces, un fichier issu de Paloma et un fichier de réconciliation DGS. Ces différents fichiers doivent être spécifiés en amont des fichiers fasta et d'annotation avec ces arguments :

- arbre des espèces : `--species_tree`
- arbre des gènes : `--gene_tree`
- fichier Paloma : `--plma_file`
- fichier de réconciliation : `--reconc_domains`

2.1.2 Modification des paramètres

Lors du développement de l'outil PhyloCharMod, les paramètres choisis pour le programme Paloma sont les paramètres conseillés pour l'identification de domaines (Onglet "help" [16]), ceux-ci semblant satisfaisants pour l'identification de modules sur les protéines ADAMTS-TSL de la matrice extracellulaire.
Les paramètres sont les suivants (les valeurs choisies sont ci-dessous entre parenthèses) :

- `-q` correspond au nombre de séquences minimum qui doivent être présentes dans un groupe (2)
- `-M` correspond à la taille maximum des alignements locaux conservés (20)
- `-t` seuil utilisé pour Dialign (10)

Au cours de mon étude sur les fibulines, la question de la robustesse de ces paramètres a été posée. Pour répondre à cette question, les paramètres t et M ont été modifiés comme suit :

- `-M` : 20, 17 et 15
- `-t` : 5, 7, 15

Les variations appliquées aux paramètres :

- `M` permettent de considérer des alignements locaux de taille différente.
- `t` permettent de filtrer les alignements locaux pour ne garder que ceux présentant une certaine similarité.

L'outil PhyloCharMod ne permet pas de changer les paramètres de Paloma au niveau du lancement du programme principal. J'ai lancé Paloma depuis le terminal du conteneur Singularity avec cette commande :

```
paloma-D -i fichier_fasta -q 2 -m 1 -M 15 -t 10 --oplma
```

Le fichier avec l'extension .oplma obtenu avec cette commande est ensuite utilisé comme entrée du programme principal de PhyloCharMod avec l'option `--plma_file`.

2.1.3 Distribution

L'installation de PhyloCharMod est possible via l'utilisation d'un conteneur *Docker* [28] avec la commande : `docker pull ghcr.io/ocmalde/phylocharmod:0.1`. La mise à disposition de l'outil sous la forme d'un conteneur permet de regrouper le code et les dépendances d'un logiciel dans une unité unique, facilement transférable d'un environnement à l'autre. Cependant, l'utilisation de Docker n'est pas toujours autorisée sur les clusters de calcul comme Genouest. Une autre solution de *conteneurisation* existe, *Singularity* [29]. Un des principaux avantages des conteneurs Singularity par rapport à la solution Docker est qu'ils sont conçus pour fonctionner en tant qu'utilisateur non privilégié, ce qui constitue un avantage certain en termes de sécurité. Par ailleurs, Singularity utilise des signatures cryptographiques, un format d'image de conteneur immuable connu sous le nom de Singularity Image Format, et un décryptage en mémoire. La plateforme de bio-informatique Genouest qui met à disposition les clusters de calcul dont nous avons besoin pour cette étude, utilise des conteneurs Singularity qui permettent normalement d'importer des images Docker préexistantes. Au cours de notre étude, l'image Docker contenant le pipeline Phylocharmod et qui était mise à disposition sur un GitHub s'est révélée incompatible avec Singularity. Plus précisément, certaines dépendances (Paloma, TreeFix) ne sont pas installées lors du lancement du programme principal de PlyloCharMod. Il fallait réaliser une installation manuelle de ces dépendances, ce qui diminue l'intérêt d'utiliser cette image Docker. Il a donc été décidé par la suite de créer un conteneur Singularity avec les codes source du logiciel et en y téléchargeant les différentes dépendances.

2.1.4 Recherche des séquences des fibulines

Outre le pipeline principal permettant de caractériser des motifs fonctionnels au sein de protéines, le conteneur possède des scripts permettant, à partir des identifiants de RefSeq des protéines humaines, d'obtenir les séquences orthologues des séquences humaines chez huit autres espèces :

- *10090 - Mus musculus*
- *9913 - Bos taurus*
- *9031 - Gallus gallus*
- *8364 - Xenopus tropicalis*
- *7955 - Danio rerio*
- *7719 - Ciona intestinalis*
- *7227 - Drosophila melanogaster*

- *6239 - Caenorhabditis elegans*

Ces huit autres espèces ont été sélectionnées au cours du développement de PhyloCHarMod comme espèce représentant l'évolution des protéines [17].
À partir des protéomes de ces huit espèces et avec le protéome humain disponible en 2023, une base de données a été réalisée avec *OrthoFinder* [30, 31]. Cette base de données contient des *orthogroups* regroupant des séquences orthologues, paralogues ainsi que les isoformes des séquences composant les protéomes. Il est donc possible, à partir des Refseq d'une protéine, d'obtenir les autres séquences présentes dans l'orthogroup de la séquence donnée en entrée. Un filtre est ensuite appliqué pour ne garder que les isoformes les plus grandes pour chaque séquence.
J'ai utilisé les scripts mis à disposition. J'ai pour cela récupéré les identifiants RefSeq de toutes les isoformes des fibulines humaines et du groupe externe en utilisant le site *Uniprot* [32]. Ces identifiants sont fournis en entrée du pipeline de manière à obtenir les séquences homologues à ces protéines chez les huit autres espèces.

2.2 Construction d'un arbre phylogénétique des fibulines

Il est possible de donner en entrée du pipeline un arbre des gènes préalablement calculé. Pour réaliser cet arbre, j'ai utilisé le script `gene_phylo.py` du pipeline. Ce dernier prend un fichier fasta comprenant les séquences ainsi que l'arbre des espèces. Cet arbre est lui-même réalisable avec le script `species_phylo.py`. Le script réalise un alignement de séquences multiples (MSA) avec l'outil *Muscle* [24]. À partir de cet alignement, les sites ayant des GAPs sont filtrés avec l'outil Trimal [33]. L'inférence phylogénétique est alors réalisée avec l'outil *PhyML* [25]. L'arbre obtenu est alors corrigé en fonction de l'arbre des espèces en utilisant l'outil *TreeFix* [26]. L'arbre obtenu est au format Newick, un format utilisé en biologie et en génétique pour décrire les relations phylogénétiques entre espèces ou taxons biologiques.

2.2.1 Enracinement de l'arbre, identification d'un groupe externe

Il est nécessaire d'identifier un groupe externe représentant des séquences extérieures à notre famille d'intérêt, mais suffisamment proches pour permettre l'enracinement. Ce groupe est appelé outGroup et permet de déterminer la branche la plus ancienne au sein de l'arbre.
Pour déterminer cet outgroup, l'ensemble des séquences canoniques des protéines de la famille des fibulines humaines ont été récupérées sur Uniprot. Un blast a été réalisé pour chaque fibuline humaine, les résultats récupérés ont permis d'identifier deux familles de protéines présentes dans chacun des résultats du blast de chaque fibuline humaine. L'outil *OrthoInspector* [34] a permis d'identifier les séquences homologues du groupe externe ainsi que celles des fibulines chez les huit autres espèces. Un arbre est alors réalisé avec l'outil NGPhylogeny [35] et une visualisation est réalisée avec l'outil en ligne, iTOL (Interactive Tree Of Life) [36].
L'arbre des gènes est alors réalisé comme décrit précédemment. Une fois l'arbre réalisé, ce dernier est visualisé grâce à l'outil iTOL. Ce site permet notamment d'enraciner l'arbre en positionnant la racine entre les groupes externes et le reste des séquences. Une fois l'arbre enraciné, le groupe externe est retiré de l'arbre en utilisant les fonctions de l'outil iTOL et ce dernier est exporté au format Newick. Les séquences correspondantes au groupe externe sont aussi retirées du fichier fasta contenant notre jeu de séquences. Les deux fichiers, le fichier fasta et celui de l'arbre enraciné, sont alors donnés en entrée au pipeline. L'arbre des gènes est spécifié avec l'option `--gene_tree`.

2.3 Fonction prédite : les Interaction protéine-protéine

Les phénotypes considérés pour cette étude sont les interactions protéine-protéine (PPI). Pour récupérer l'ensemble des interactions protéine-protéine, nous avons utilisé l'outil *PSICQUIC* [37]. Cet outil permet un accès aux informations d'interactions moléculaires présentes sur différentes bases de données. Pour chaque membre de la famille des fibulines humaines, une recherche est réalisée sur le site PSICQUIC view et les données sont alors téléchargées au format CSV. Les informations obtenues sont triées pour ne conserver que les PPI uniques impliquant les protéines humaines (>90%) et le graphe d'interaction des protéines est visualisé à l'aide de l'application Cytoscape [38]. Seules les interactions protéine-protéine humaine sont considérées, étant donné que les données expérimentales concernant les autres espèces sont rares.

3 Résultat

3.1 Installation de PhyloCharMod dans un conteneur singularity

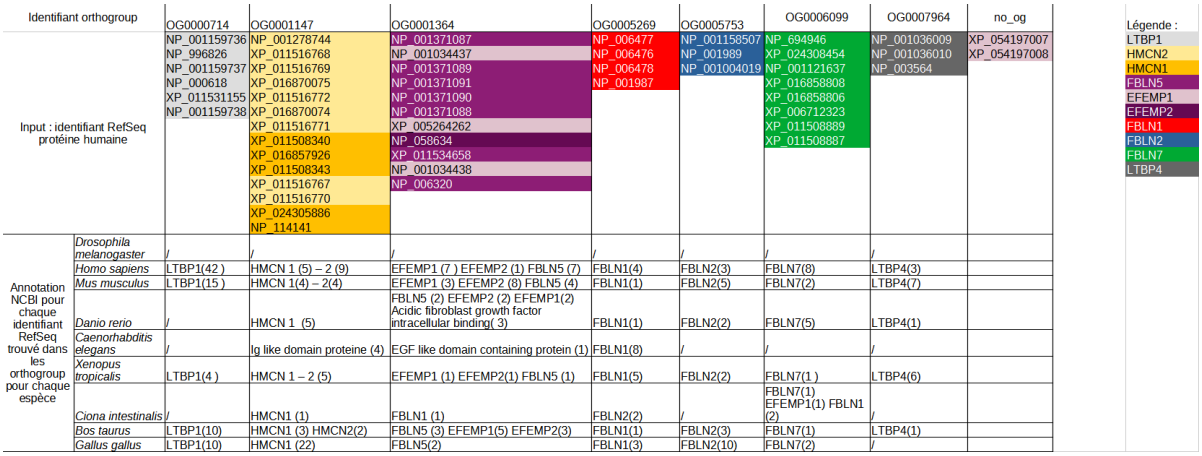
Comme nous l'avons détaillé dans la section Distribution (partie 2.1.3), l'image Doker fournie par Olivier Dennler s'est révélée incompatible avec l'outil Singularity utilisé par la plateforme de bio-informatique Genouest. Brièvement, nous avons créé un conteneur Singularity avec les codes source du logiciel, mis à jour les versionnages de packages et téléchargé les différentes dépendances, notamment celles liées à l'outil Paloma.

3.2 Identification du groupe externe

La recherche au sein des fichiers issus du blast réalisé avec les séquences protéiques de fibulines humaines, a permis d'identifier deux familles de protéines proches des fibulines : les LTBP (Latent-transforming growth factor beta-binding protein) et les FBLN (Fibrillins). La réalisation de l'arbre des fibulines et ces familles de protéines a permis de mettre en évidence que la famille des LTBP est plus proche des fibulines que les fibrilins. Les membres de la famille des LTBP -1 et -4 constitueront donc notre groupe externe.

3.3 Construction d'un jeu de données des séquences

Comme expliqué dans la partie 2.1.4, la construction du jeu de données des séquences a été réalisée à partir de la base de données mise en place par Oliver Dennler avec l'outil OrthoFinder et disponible au sein du conteneur. L'interrogation avec les identifiants RefSeq des séquences des protéines humaines des huit fibulines et des LTBP a permis de sélectionner huit orthogroups chacun nommé par un identifiant unique propre à la base de données. Parmi ces huit groupes, cinq correspondent à des groupes contenant des fibulines, deux correspondent au groupe extérieur constitué des LTBP1 et LTBP4, enfin un dernier groupe contient deux identifiants RefSeq fibuline-3 humaine, ces derniers ne sont pas présents au sein des orthogroups présents dans la base de données. Concernant les cinq groupes contenant des fibulines: le groupe *OG0000714* comprend 65 séquences et regroupe les hémicentines-1 et -2 humaines. Le groupe *OG0001364* contient 59 séquences et regroupe les fibulines-3,-4 et -5. Enfin, les groupes *OG0005269*, *OG0005753* et *OG0006099* contiennent respectivement les fibulines-1, -2 et -7. *Drosophila melanogaster* ne présente aucune séquence dans ces orthogroups (Figure 5).



Identifiant orthogroup		OG0000714	OG0001147	OG0001364	OG0005269	OG0005753	OG0006099	OG0007964	no_og
Input : identifiant RefSeq protéine humaine		NP_001159736 NP_996826 NP_001159737 NP_000618 XP_011531155 NP_001159738	NP_001278744 XP_011516768 XP_011516769 XP_016870075 XP_011516772 XP_016870074 XP_011516771 XP_011508340 XP_016857926 XP_011508343 XP_011516767 XP_011516770 XP_024305886 NP_114141	NP_001371087 NP_001034437 NP_001371089 NP_001371091 NP_001371090 NP_001371088 XP_005264262 NP_058634 XP_011534658 NP_001034438 NP_006320	NP_006477 NP_006476 NP_006478 NP_001987	NP_001158507 NP_001989 NP_001004019	NP_694946 XP_024308454 NP_001121637 XP_016858808 XP_016858806 XP_006712323 XP_011508889 XP_011508887	NP_001036009 NP_001036010 NP_003564	XP_054197007 XP_054197008
Annotation NCBI pour chaque identifiant RefSeq trouvé dans les orthogroup pour chaque espèce	*Drosophila melanogaster*	/	/	/	/	/	/	/	
	Homo sapiens	LTBP1(42)	HMCN 1 (5) – 2 (9)	EFEMP1 (7) EFEMP2 (1) FBLN5 (7)	FBLN1(4)	FBLN2(3)	FBLN7(8)	LTBP4(3)	
	Mus musculus	LTBP1(15)	HMCN 1(4) – 2(4)	EFEMP1 (3) EFEMP2 (8) FBLN5 (4)	FBLN1(1)	FBLN2(5)	FBLN7(2)	LTBP4(7)	
	Danio rerio	/	HMCN 1 (5)	FBLN5 (2) EFEMP2 (2) EFEMP1(2) Acidic fibroblast growth factor intracellular binding(3)	FBLN1(1)	FBLN2(2)	FBLN7(5)	LTBP4(1)	
	Caenorhabditis elegans	/	Ig like domain proteine (4)	EGF like domain containing protein (1)	FBLN1(8)	/	/	/	
	Xenopus tropicalis	LTBP1(4)	HMCN 1 – 2 (5)	EFEMP1 (1) EFEMP2(1) FBLN5 (1)	FBLN1(5)	FBLN2(2)	FBLN7(1)	LTBP4(6)	
	Ciona intestinalis	/	HMCN1 (1)	FBLN1 (1)	FBLN2(2)	/	FBLN7(1) EFEMP1(1) FBLN1 (2)	/	
	Bos taurus	LTBP1(10)	HMCN1 (3) HMCN2(2)	FBLN5 (3) EFEMP1(5) EFEMP2(3)	FBLN1(1)	FBLN2(3)	FBLN7(1)	LTBP4(1)	
	Gallus gallus	LTBP1(10)	HMCN1 (22)	FBLN5(2)	FBLN1(3)	FBLN2(10)	FBLN7(2)	/	

Légende :
LTBP1
HMCN2
HMCN1
FBLN5
EFEMP1
EFEMP2
FBLN1
FBLN2
FBLN7
LTBP4

Figure 5. **Tableau récapitulatif des informations issues de l'utilisation des données Orthofinder de l'outil.** Chaque colonne commençant par OG correspond à un orthogroup spécifique à la base de données. La première ligne permet de connaitre l'identifiant de l'orthogroup au sein de la base de données présente dans l'outil. La deuxième ligne (Input) correspond au identifiant RefSeq de l'ensemble des isoformes des protéines données en entrée. Les couleurs indiquent à quelle protéine correspondes chaque RefSeq donné, voir légende à droite. Le bloc de ligne suivant permet pour chaque espèce de connaître les annotations NCBI de chaque RefSeq composant l'orthogroup et le nombre de séquences par annotation. Légende : LTBP : Latent-transforming growth factor beta-binding protein, FBLN : fibuline, HMNC : hémicentine, EFEMP1 : fibuline-3, EFEMP2 : fibuline-4

Parmi les séquences présentes dans les orthogroups, nous pouvons remarquer que l'annotation présente dans la base NCBI diffère de celle des séquences orthologues humaines présentes dans ce groupe. Par exemple, dans l'orthogroup OG0006099 qui correspond au groupe de la fibuline-7 humaine présente pour l'espèce *Ciona intestinalis* des séquences annotées fibuline-1 et fibuline-3 (EFEMP1).

Après la filtration réalisée pour ne garder que les isoformes les plus longues, le nombre de séquences gardées par espèce est récapitulé dans la table 1.

Table 1. Nombre de séquences par espèce dans le jeu de données après obtention des séquences homologues et de la sélection des isoformes les plus longues

Espèces	**Nombre de séquences**
Homo sapiens	10
Mus musculus	10
Bos taurus	10
Gallus gallus	8
Xenopus tropicalis	12
Danio rerio	13
Ciona intestinalis	5
Drosophila melanogaster	0
Caenorhabditis elegans	2

Parmi les séquences gardées, la séquence (RefSeq : XP_031754339.1) annotée comme une fibuline-1 de *Xenopus tropicalis*, présente une longueur de 116 acides aminés. Cette protéine est bien plus courte que toutes les autres fibulines de l'orthogroup (703 acides aminés pour la

fibuline-1 humaine). Cette séquence a donc été supprimée pour la suite de cette étude. Le jeu de séquences final contient 59 séquences correspondant à des fibulines et dix séquences correspondant au groupe externe (LTBP).

3.4 Construction de la phylogenie des Fibulines

L'arbre obtenu à partir des scripts présents dans PhyloCharMod (Figure 6) montre une topologie similaire à celle précédemment décrite par [39] avec un regroupement des hémicentines-1 et -2 d'une part et un regroupement des fibulines-1 et -2 d'autre part, les fibulines-3, -4 et -5 restants toutes les trois proches au sein de l'arbre. L'arbre contient 116 gènes ancestraux. Par la suite, chaque gène ancestral sera nommé par un G suivi d'un nombre propre à chaque gène.

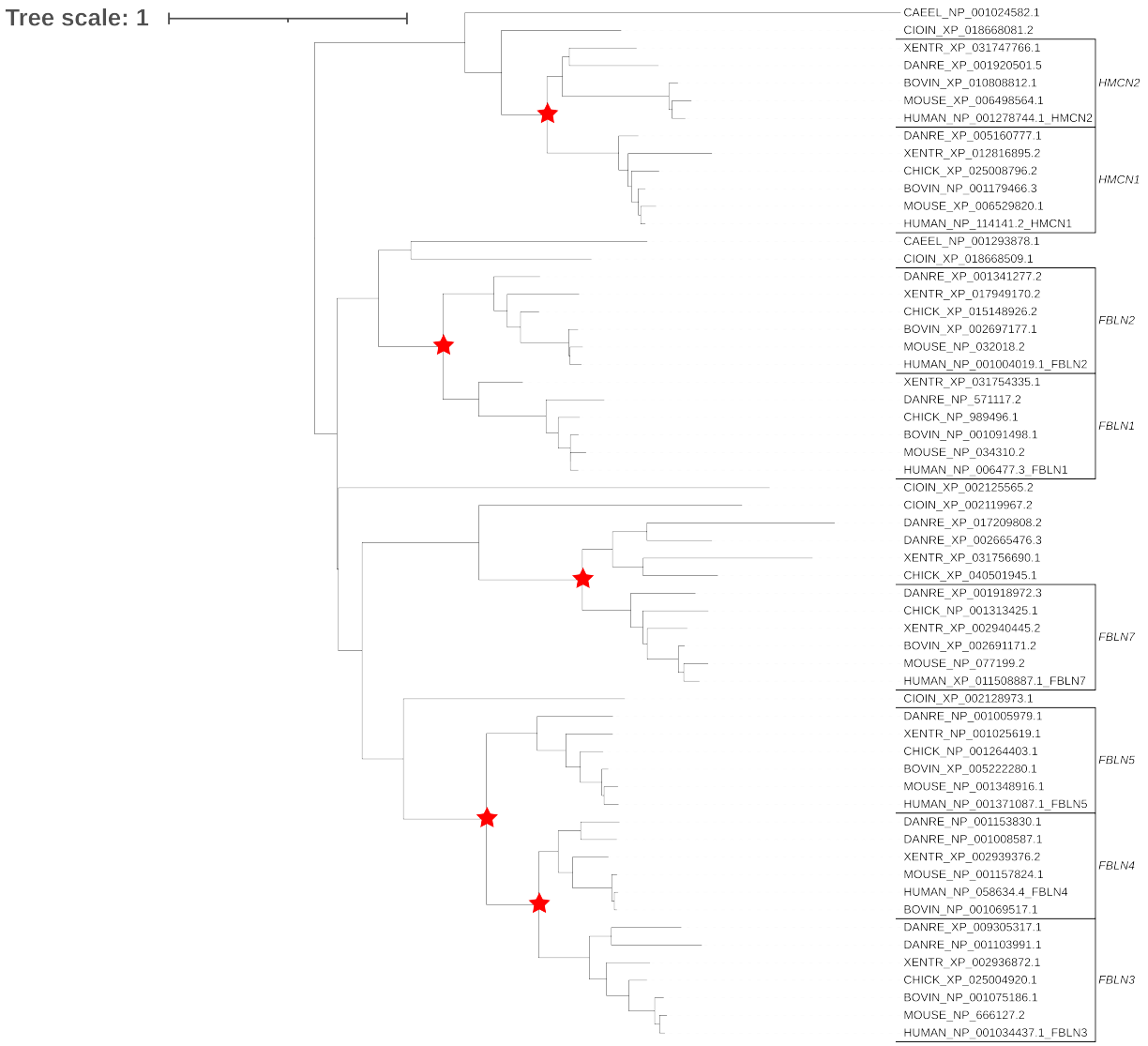


Figure 6. **Arbre des gènes des fibulines obtenu après enracinement de l'arbre et suppression du groupe externe pour les fibulines** (figure produite avec l'outil iTOL). Les feuilles de l'arbre sont annotées suivant le modèle : espèce_RefSeq, si la protéine correspond à une fibuline humaine, leur annotation sera indiquée. Les différents sous-arbres indiqués à droite correspondent aux orthologues des fibulines humaines (chaque nœud du sous-arbre est un nœud de spéciation). Les étoiles rouges correspondent à des duplications identifiées. La barre d'échelle en haut à gauche indique la longueur des branches, en espérance du nombre de substitutions par site. *Légende :* CAELL : *Caenorhabditis elegans*, CION : *Ciona intestinalis*,DANRE : *Danio rerio*, XENTR : *Xenopus tropicalis*, CHICK : *Gallus gallus*, BOVIN : *Bos taurus*, MOUSE : *Mus musculus*, HUMAN : *Homo sapiens*, HMCN1 : hémicentine-1, HMCN2 : hémicentine-2, FBLN2 : fibuline-2, FBLN1 : fibuline-1, FBLN7 : fibuline-7, FBLN5 : fibuline-5, FBLN4 : fibuline-4 (EFEMP2) et FBLN3 : fibuline-3 (EFEMP1)

L'arbre des gènes construit ici montre des événements de duplication (Figure 6) au niveau de l'ancêtre des *Euteleostomi* (vertébrés à mâchoire, incluant *Danio* et *Homo*) (arbre des espèces disponibles en annexe 1) pour les différents regroupements : les hémicentines-1 et -2, les fibulines-1 et -2 et les fibulines -3-4 et -5.
Le sous-arbre des fibulines-7 montre aussi une duplication (Figure 6) au niveau de l'ancêtre des *Chordata* cependant l'une de ces copies n'est présente que chez les genres *Danio*, *Xenopus* et *Gallus*, ceci suggère une perte de cette copie chez l'ancêtre des mammifères (*Boreoeutheria*).

De plus, nous observons pour chaque fibuline plusieurs copies chez *Danio rerio*, ce qui est cohérent avec la duplication connue de ce génome [40].

3.5 Identification des modules par l'outil Paloma et comparaison avec les domaines de type Pfam

La figure 7 permet de montrer la distribution en modules le long des séquences protéiques des fibulines humaines. Les modules sont répartis tout le long des séquences des fibulines humaines, hormis pour l'hémicentine-2 où aucun module n'a été identifié au niveau des domaines Immunoglobuline-like répétés (visibles en gris). La figure 2 montre l'existence d'un "fibuline like module" partagé par l'ensemble des fibulines [2]. Un tel domaine serait normalement visible sous la forme d'un module conservé chez l'ensemble des fibulines. Cependant, aucun module n'est identifié en position C-terminale chez les huit fibulines humaines. Par contre, trois fibulines (fibuline-3, -4 et -5) partagent un module en position C-terminale (module vert hexagonal, figure 7).

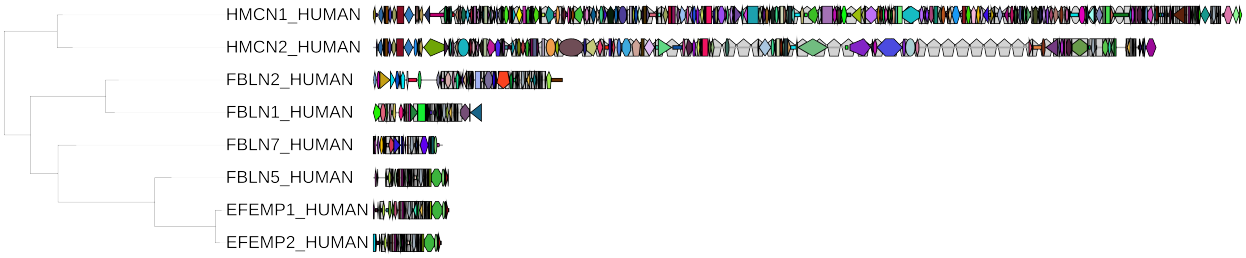


Figure 7. Distribution des modules identifiés par l'outil Paloma dans les séquences des protéines des fibulines humaines. L'arbre des fibulines humaines a été obtenu à partir de l'arbre des gènes de référence des fibulines montré dans la figure 6. Les modules sont représentés par une combinaison unique de formes et de couleurs sur chaque séquence protéique humaine avec l'outil Itol. Les formes grises correspondent à la composition en domaine de chaque séquence. Légende : HMCN1 : hémicentine-1,HMCN2 : hémicentine-2, FBLN2 : fibuline-2, FBLN1 : fibuline-1, FBLN7 : fibuline-7, FBLN5 : fibuline-5, EFEMP2 : fibuline-4 et EFEMP1 : fibuline-3

3.6 Identification des réseaux d'interactions des protéines humaines

La recherche des interactions protéine-protéine (PPI) a permis d'identifier 483 interactions impliquant au moins une des huit fibulines humaines (Figure 8). Pour l'étude, seules les PPI partagées (Figure 9) par au moins deux fibulines sont considérées. Sans entrer dans le détail des protéines interagissant avec ces fibulines, nous pouvons noter que 14 protéines interagissent avec les fibulines-1,-2,-3,-4,-5 et de façon remarquable, 10 d'entre elles sont des protéines de la matrice extracellulaire : ELN, EMILIN1, FBN1, FBN2, MFAP1, MFAP2, MFAP3, MFAP4, MFAP5 et VTN. Les fibulines étant des protéines sécrétées, leur interaction avec d'autres constituants de la matrice extracellulaire n'est pas surprenante. À l'inverse, ces interactions ne sont pas partagées par les fibulines-6, -7 et -8 et pourraient s'expliquer par leurs études plus récentes et donc par un manque de données et/ou par des fonctions différentes. Ainsi la fibuline-7 a été découverte récemment en 2007 [41] et présente des spécificités de distribution et de fonctions [42]. Les fibulines-6 et -8 correspondent aux hémicentines initialement identifiées chez *Caenorhabditis elegans* mais dont la fonction reste encore très peu documentée [43]. Nous observons également cinq protéines partagées par les fibulines -1 et -2, les plus étudiées à ce jour.

Parmi ces cinq protéines, ACAN, VCAN et HSPG2 sont trois protéoglycanes partageant des caractéristiques biochimiques spécifiques et que l'on retrouve plus particulièrement dans les tissus cartilagineux [44]. L'ensemble des intersections des PPI est disponible dans l'annexe 2.

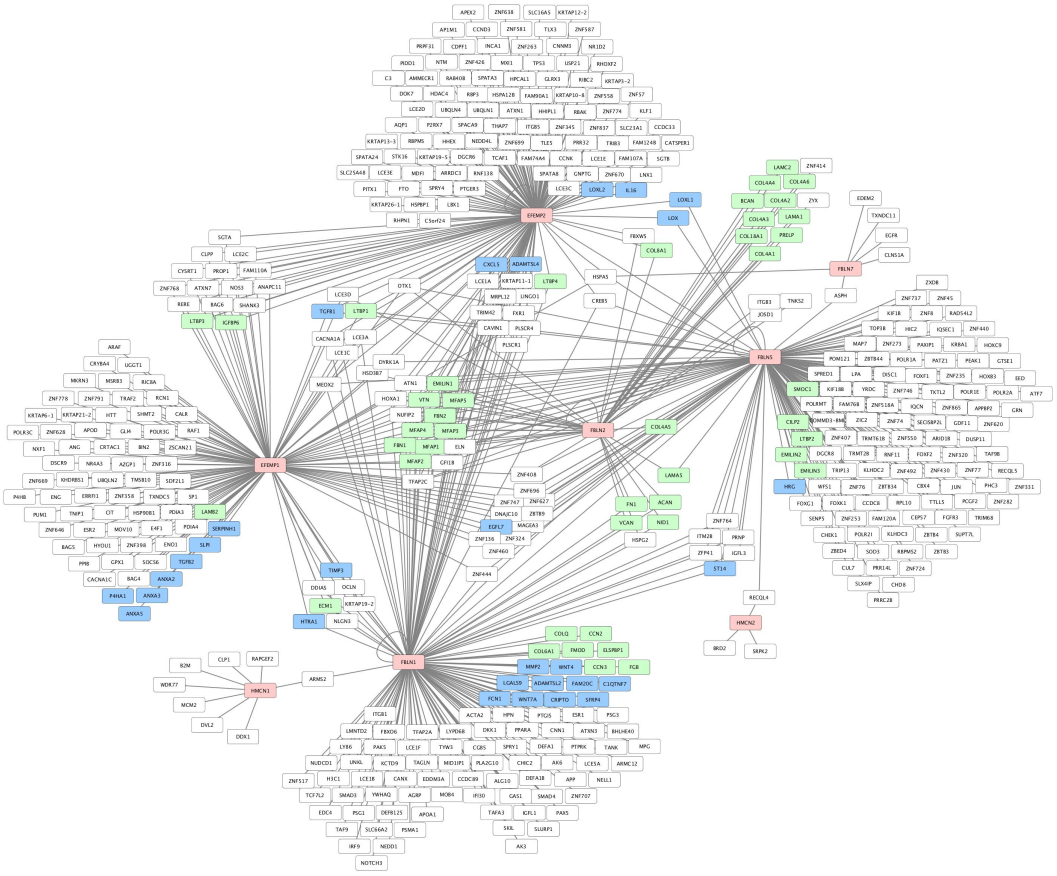


Figure 8. **Réseau d'interaction protéine-protéine des fibulines humaines.** La visualisation a été réalisée avec Cytoscape [38] et montre les 483 PPI des fibulines humaines. Les nœuds rouges représentent les fibulines, les nœuds bleus sont des protéines associées aux protéines extracellulaires, les nœuds verts sont les protéines du core matrisome et les nœuds blanc aux autres protéines interagissant avec les fibulines.

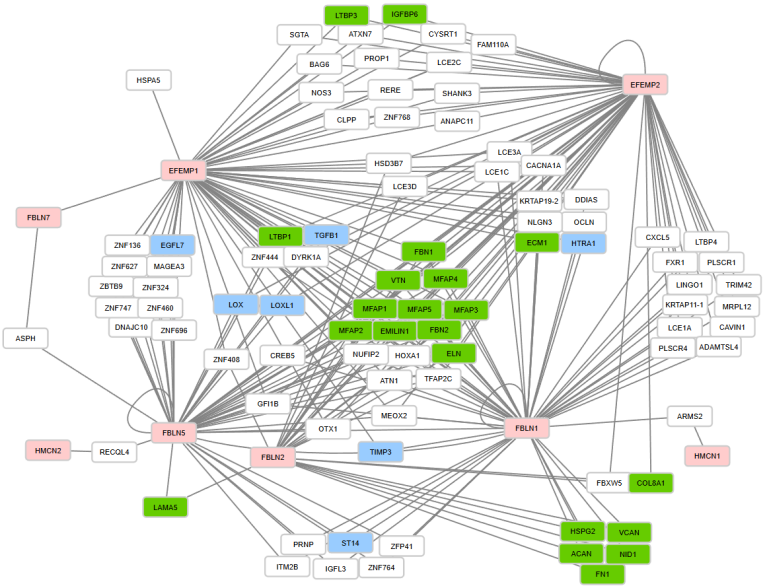


Figure 9. **Réseau des interactions protéine-protéine partagées par au moins deux fibulines humaines.** La visualisation a été réalisée avec l'outil Cytoscape [38] et montre les 97 PPI partagées par au moins deux fibulines humaines. Les nœuds rouges représentent les fibulines, les nœuds bleus sont des protéines associées aux protéines extracellulaires, les nœuds verts sont les protéines du core matrisome et les nœuds blancs aux autres protéines interagissant avec les fibulines.

3.7 Mise en évidence des événements de Co-apparition de PPI et de modules conservés

3.7.1 Reconstruction de l'histoire des modules avec SeadogMD

Comme expliqué dans la partie 1.3.2, l'histoire évolutive des modules a été réalisée avec SEADOG-MD. Un total de 508 modules conservés ont été identifiés par Paloma. Parmi les 116 gènes ancestraux qui composent l'arbre des fibulines, 48 présentent un gain de modules.
La présence ou non de modules chez différentes fibulines permet d'estimer l'ancêtre depuis lequel chaque module est conservé. Le module vert en C-terminale présent chez les fibuline-3, -4 et -5 en est un bon exemple (figure 10) : ce dernier semble apparaître au niveau de l'ancêtre *Euteleostomi*. Le module bleu présent chez fibuline-7 en C-terminale paraît lui aussi être apparu à l'ancêtre *Euteleostomi*. Trois modules sont eux retrouvés à l'extrémité C-terminale des hémicentines-1, ces modules apparaissent chez les descendants des *Eustelostomi*. Ces modules sont présents spécifiquement chez les vertébrés du groupe Euteleostomi, incluant selon nos données les genres *Danio*, *Xenopus*, *Gallus*, *Bos*, *Mus* et *Homo*.

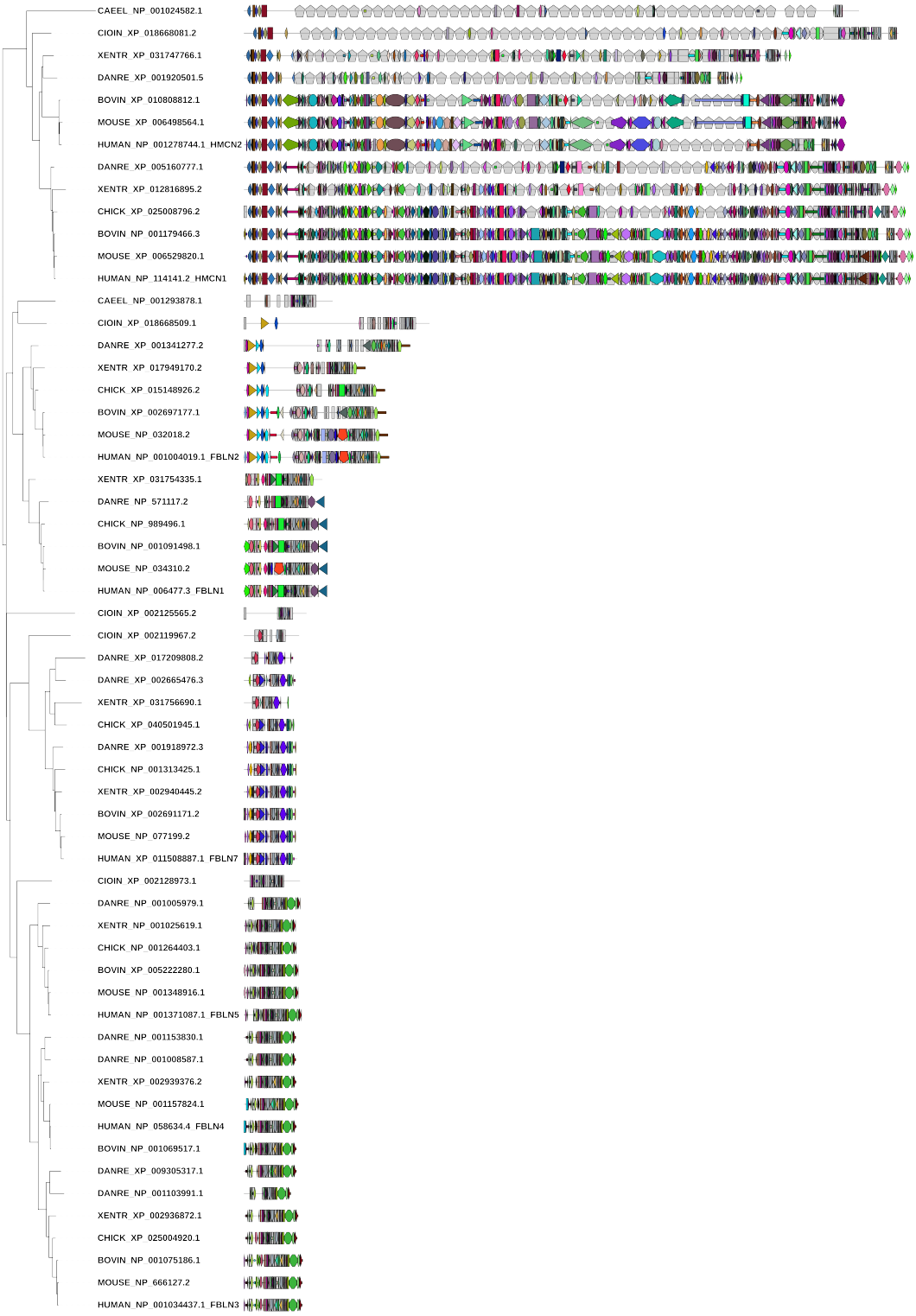


Figure 10. **Arbre des gènes des fibulines avec la composition en modules de l'ensemble des séquences protéiques des fibulines pour l'ensemble des espèces.** Chaque module est représenté par une combinaison de forme et de couleur unique. *Légende :* CAELL : *Caenorhabditis elegans*, CION : *Ciona intestinalis*,DANRE : *Danio rerio*, XENTR : *Xenopus tropicalis*, CHICK : *Gallus gallus*, BOVIN : *Bos taurus*, MOUSE : *Mus musculus*, HUMAN : *Homo sapiens*, HMCN1 : hémicentine-1, HMCN2 : hémicentine-2, FBLN2 : fibuline-2, FBLN1 : fibuline-1, FBLN7 : fibuline-7, FBLN5 : fibuline-5, FBLN4 : fibuline-4 (EFEMP2) et FBLN3 : fibuline-3 (EFEMP1)

3.7.2 Reconstruction de l'histoire des interactions protéine-protéine avec PAstml

La reconstruction des scénarios ancestraux pour les PPI est visible sur la figure 11. Cette représentation nous permet de voir que les 14 PPI ayant interagi avec les cinq fibulines remontent dans l'arbre jusqu'au nœud séparant les hémicentines des autres fibulines. Ces interactions sont estimées exister depuis l'ancêtre *Bilateria*. Les PPI comme celles présentes aux niveaux de la fibuline-7, de l'hémicentine-1 et de l'hémicentine-2 humaines ne remontent pas dans l'arbre malgré une interaction commune avec d'autres fibulines.

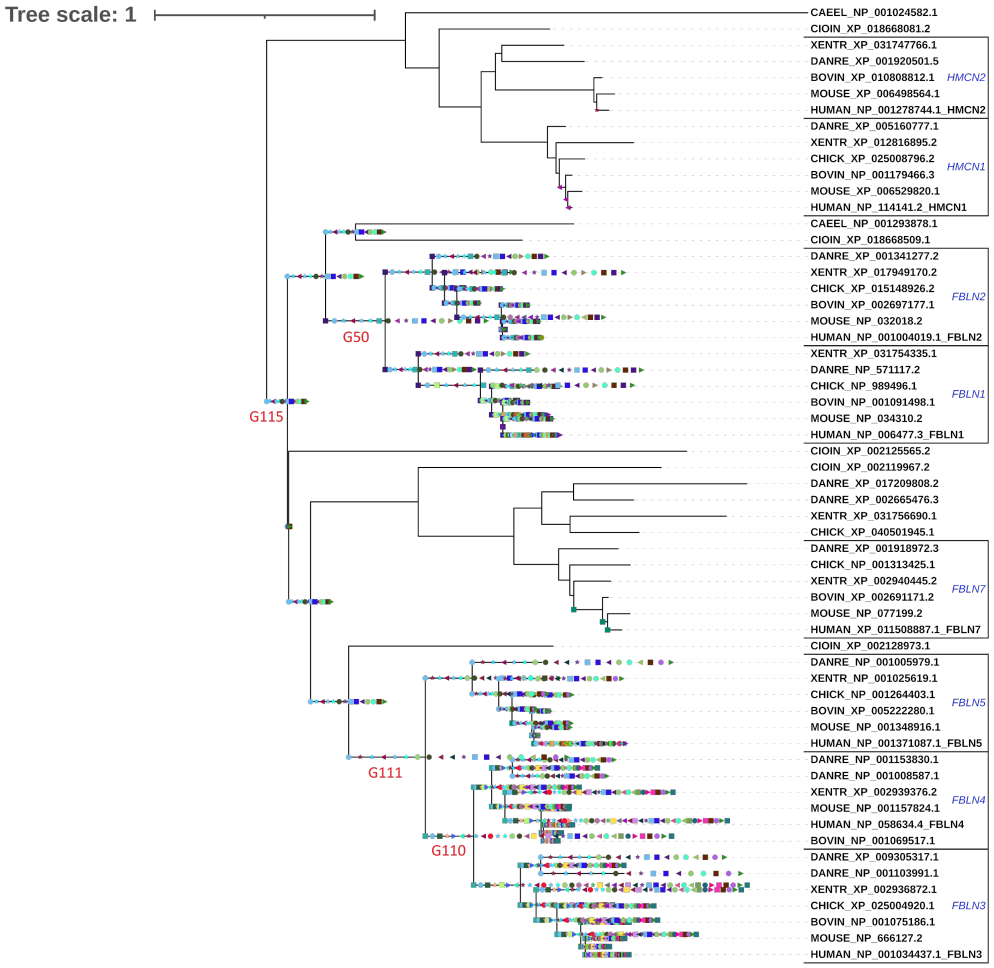


Figure 11. **Arbre des gènes des fibulines avec la visualisation des scénarios ancestraux.** Chaque combinaison unique de forme et de couleur représente une des 97 PPI partagées par au moins deux fibulines. On peut noter une disparité de la répartition des interactions entre le sous-groupe fibulines -3, -4 et -5 et le sous-groupe fibulines-1 et -2. À l'inverse, les hémicentines et la fibuline-7 sont annotées avec une seule interaction. Le nom des gènes ancestraux présentant une co-apparition module/annotation avec deux descendants humains ou plus est indiqué en rouge. *Légende :* CAELL : *Caenorhabditis elegans*, CION : *Ciona intestinalis*,DANRE : *Danio rerio*, XENTR : *Xenopus tropicalis*, CHICK : *Gallus gallus*, BOVIN : *Bos taurus*, MOUSE : *Mus musculus*, HUMAN : *Homo sapiens*, HMCN1 : hémicenyine-1, HMCN2 : hémicentine-2, FBLN2 : fibuline-2, FBLN1 : fibuline-1, FBLN7 : fibuline-7, FBLN5 : fibuline-5, FBLN4 : fibuline-4 (EFEMP2) et FBLN3 : fibuline-3 (EFEMP1).

3.7.3 Identification de 13 nœuds de co-apparitions

L'étude des nœuds de l'arbre gènes permet d'identifier des gènes ancestraux où il y a co-apparition de modules conservés et d'interactions protéine-protéine permettant d'associer des modules de séquences potentiellement impliqués dans ces interactions. L'analyse réalisée avec les paramètres par défaut de PhyloCharMod a permis de mettre en évidence 13 nœuds ancestraux (Table 2) dont quatre d'entre eux (G50, G110, G111, G115) présentent deux descendants humains ou plus (localisation dans l'arbre voir figure 11). Au cours de ce travail, nous avons exploré plus précisément les événements de co-évolution au niveau de ces quatre gènes ancestraux.

Table 2. Table des 13 nœuds présentant les 15 noeud de co-apparitions module-PPI. Chaque nœud est défini par son nom (Gene), les Descendants, les PPI et les modules. Les nœuds G50, G110, G111 et G115 présentent au moins deux séquences humaines comme descendants.

Gene	**Descendant**	**PPI**	**Module**
G18	Hémicentine 1	ARMS2	B374 B365 B338 B502 B269 B484 B492 B457 B441 B372 B435 B465 B439 B419 B463 B447 B290 B413 B482 B453 B451 B282 B444 B268 B327 B418
G35	FBLN1	ARMS2 CREB5 FBLN1 FBLN5 IGFL3 ITM2B PRNP ST14 ZFP41 ZNF444 ZNF764	B657 B653 B681 B615
G37	FBLN1	CACNA1A GFI1B LCE1C LCE3A MEOX2	B686 B677 B684
G46	FBLN2	HSD3B7 LAMA5 LCE3D	B705 B714 B880 B703 B615 B701 B687
G47	FBLN2	OTX1	B697 B715 B713
G50	FBLN1, 2	ACAN FN1 HSPG2 NID1 VCAN	B741 B736 B730 B737 B680 B663 B674 B610 B739 B679 B728
G68	FBLN7	ASPH	B746 B745
G82	FBLN5	ASPH	B878
G94	FBLN4	CREB5 FBLN5	B651 B615 B648
G109	FBLN3	CACNA1A HSD3B7 LCE1C LCE3A LCE3D	B244
G110	FBLN3, 4	ANAPC11 ATXN7 BAG6 CLPP CYSRT1 FAM110A IGFBP6 LCE2C LTBP3 NOS3 PROP1 RERE SGTA SHANK3 ZNF768	B629 B617 B593
G111	FBLN3, 4, 5	DYRK1A GFI1B LTBP1 MEOX2 OTX1 TGFB1	B626 B625 B612 B644 B622 B624 B595 B610 B628
G115	FBLN1, 2, 3, 4, 5, 7	ATN1 ELN EMILIN1 FBN1 FBN2 HOXA1 MFAP1 MFAP2 MFAP3 MFAP4 MFAP5 NUFIP2 TFAP2C VTN	B643 B883 B56 B769

Le gène le plus ancestral de l'arbre, G115 présente un gain de quatre modules (B643, B883, B56 et B769) dont un (B643) est présent chez trois de ses descendants humains (fibuline-2, -4, -5). Ces modules sont donc conservés depuis la racine de l'arbre, ce qui est interprété comme le signe d'une forte sélection naturelle et d'un phénotype important commun aux fibulines. Le module B56 est présent seulement chez la fibuline-1 et le module B769 n'est présent que chez la fibuline-7. La fibuline-3 ne présente aucun de ces modules, suggérant une perte de modules. 14 PPI sont remontées au niveau de ce gène ancestral dont 10 sont des protéines de la matrice extracellulaire interagissant avec les fibulines-1,-2, -3, -4, -5 (voir table 2).

Le gène ancestral G50, ayant parmi ses descendants les fibulines-1 et -2, a connu un gain de deux modules communs chez les deux fibulines humaines : B680 et B663. Ces deux modules conservés ont co-évolué avec les cinq PPI interagissant avec deux protéines, à savoir ACAN FN1 HSPG2 NID1 VCAN.

Le gène ancestral G111 présente un gain de neuf modules (B626, B625, B612, B644, B622, B624, B595 et B610 B628). Parmi ces modules, huit sont communs aux fibulines-3,-4 et -5, le dernier module n'est lui partagé que par les fibulines-4 et -5. Nous observons cinq PPI qui remontent à ce gène ancestral G111 et impliquent les protéines TGFB1, LTBP1, DYRK1A, GFI1B, MEOX2, OTX1, cependant seuls les protéines TGFB1, LTBP1, DYRK1A sont spécifiquement partagés par les fibulines-3, -4 et -5, GFI1B, MEOX2, OTX1 étant aussi partagés avec d'autres fibulines ailleurs dans l'arbre, suggérant des processus de convergences au cours de l'évolution. Pour les trois PPI impliquant TGFB1, LTBP1, DYRK1A, il est intéressant de noter que LTBP1 est la protéine qui se lie à la forme latente du TGFB1 et que par ailleurs DYRK1A a été décrit comme un co-régulateur avec le TGFB1 de différents processus biologiques [45], [46], DYRK1A interagissant notamment avec la protéine effectrice du signal induit parle TGFB1 [47]. Pour résumer, au niveau de ce gène ancestral G111, les fibulines-3, -4 et -5 partagent huit modules et trois PPIs en lien avec les fonctions du TGFB1.

Le gène ancestral G110 présente trois modules (B629, B617 et B593) dont deux (B629 et B593) présents chez les deux fibulines humaines -3 et -4, descendantes de ce nœud et aussi dénommées EFEMP1 et EFEMP2. Par ailleurs, on note 14 PPI qui sont remontées au niveau du gène ancestral de ces deux fibulines, ces protéines n'ayant pas de relations directes entre elles. Parmi celles-ci, on note la présence de LTBP3, un autre membre de la famille des protéines fixant la forme latente du TGFB1.

3.8 Analyse de la robustesse des prédictions de co-apparition

Afin de tester la robustesse de nos prédictions sur la co-apparition de modules conservés et de PPI, nous avons fait varier les paramètres de l'outil Paloma, calculant les modules.

Table 3. Table représentant pour chaque paramètre de paloma testé le nombre de modules obtenus et les gènes présentant une co-apparition.

Paramètre	Nombre de modules	Gène avec co-apparition
t5M15	636	G18, G35, G37, G46, G47, **G50**, G68, G82, G93, G94, G96, **G110**, **G111**, **G115**
t5M17	591	G18, G35, G37, G46, G47, **G50**, G68, G82, G93, G94, G96, G109, **G110**, **G111**, **G115**
t5M20	531	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G105, G109, **G110**, **G111**, **G115**
t7M15	695	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G96, G105, G109, **G110**, **G111**, **G115**
t7M17	581	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G109, **G110**, **G111**, **G115**
t7M20	570	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G96, G109, **G110**, **G111**, **G115**
t10M15	714	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G105, G109, **G110**, **G111**, **G115**
t10M17	629	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G96, G109, **G110**, **G111**, **G115**
t10M20	508	G18, G35, G37, G46, G47, **G50**, G68, G82, G94, G109, **G110**, **G111**, **G115**

Afin de tester la robustesse de nos prédictions sur la co-apparition de modules conservés et de PPI, nous avons fait varier les paramètres. La variation des paramètres de l'outil Paloma conduit à un nombre de modules trouvés variant de 508 à 714 modules (table 3). Parmi les nœuds présentant un gain de module, 15 présentent une co-apparition avec une PPI. Les nœuds mis en jeu sont : G18, G35, G37, G46, G47, G50, G68, G82, G93, G94, G96, G109, G110, G111 et G115. Cependant, certains nœuds de co-apparition ne sont présents que pour quelques paramètres donnés comme : G93, G96, G109. Les nœuds de co-apparition présentant deux ou plus de descendants humains trouvés précédemment (section 3.7.3) sont eux retrouvés à chaque modification des paramètres de l'outil Paloma, calculant les modules.

4 Discussion

4.1 L'identification des séquences pour notre jeu de données initiales constitue l'étape clé de notre étude

Lors de la construction du jeu de séquences à partir de la base de données constituée par Olivier Dennler et issue d'une analyse par l'outil Orthofinder, je n'ai pas identifié de séquences pour l'espèce *Drosophila melanogaster*1. Des recherches sur les bases de données Uniprot et NCBI ont confirmé l'absence de séquences correspondant aux fibulines chez cette espèce. Cependant, des recherches complémentaires ont montré des séquences correspondant à des fibulines chez les espèces *Drosophila kikkawai* et *Drosophila mauritiana*. Il serait intéressant de recommencer l'analyse en incluant ces séquences protéiques en modifiant de façon manuelle le fichier fasta regroupant les séquences en entrée.

Ces observations nous ont amenées à réfléchir sur la limitation de cette base de données présente dans le container. Elle dépend des protéomes donnés en entrée à Orthofinder. À titre

d'exemple, les identifiants RefSeq XP_054197007 et XP_054197008, donnés en entrée, se trouvent dans l'othogroup no_og ce dernier signifie que l'outil ne les trouve pas dans la base de données. Ces deux séquences ont été publiées sur NCBI en octobre 2023, après la création de la base de données par Olivier Dennler et donc nécessite une mise à jour. La version de la séquence est indiquée par un chiffre après le point dans un identifiant RefSeq et pour éviter ce problème d'identification, nous avons été amenés à modifier le programme afin de ne pas prendre en compte la version de la séquence.

Un autre élément de réflexion est venu de l'identification d'une séquence de *Xenopus* avec une séquence extrêmement courte, séquence prédite sans preuves expérimentales. La majorité des séquences utilisées, hors humain, correspondent à des séquences prédites à partir de séquences génomiques par la méthode de prédiction de gène : Gnomon. Ce logiciel de prédictions de gène utilise des séquences génomiques, des transcrit ou des séquences de protéines pour réaliser des alignements entre les séquences protéiques et les transcrits sur le génome de référence afin d'identifier des régions codantes potentielles. Ces derniers sont donc prédits sur la base de l'alignement, mais aussi des caractéristiques de la séquence. Les exons prédits sont alors assemblés en gène complet.
Malgré l'automatisation dans la sélection des séquences homologues chez les différences espèces via l'utilisation des données présentes au sein du container, une vérification manuelle par l'utilisateur de l'outil reste indispensable. Cette vérification doit porter tant sur la pertinence des séquences sélectionnées, mais aussi de la compréhension de l'absence de certaines séquences comme cela a été le cas ici avec la *Drosophila melanogaster*.

4.2 Identification de gènes ancestraux présentant un gain de module caractéristique d'un sous-groupe ou un gain de fonction chez la famille des fibulines

L'arbre construit au cours de cette étude (Figure 6) montre des similarités dans sa topologie à l'arbre précédemment décrit par [39]. Les deux arbres montrent une proximité évolutive des deux hémicentines mais aussi des fibulines-1 et -2 ainsi que des fibulines-3, -4 et -5. Notre arbre a permis de mettre en évidence une duplication de la fibuline-7 au niveau de l'ancêtre des Chordata. Cependant, l'un des gènes issus de cette duplication a été perdu chez l'ancêtre des *Boreoeutheria*. L'observation de plusieurs copies de fibulines pour l'espèce *Danio rerio* est cohérente avec son histoire évolutive puisque ce dernier a eu une duplication de l'ensemble de son génome [40].

Les modules sont répartis tout le long des séquences protéiques, néanmoins l'absence de modules au niveau de certains domaines Immunoglobulin-like chez les hémicentines est à noter. Cette absence de module sur ces domaines répétés peut être due à l'alignement de ces segments à d'autres endroits des protéines. La présence de modules présents chez des sous-groupes de protéine indique la conservation de ces segments au sein de ce sous-groupe. Leur présence serait donc une signature de ce sous-groupe. Par exemple, le module B771 (losange bleu en position C-terminale pour le cluster FBLN7 figure 10)) serait un module signature des fibulines-7 ayant pour ancêtre le nœud G72 (ancêtre *Chordata*). Un autre module intéressant est retrouvé en position C-terminale (domaine fibuline-like) des protéines composant les clusters FBLN5, FBLN4 et FBLN3, il est, sur la figure 10, représenté avec un hexagone vert et nommé B624. Ce module est gagné au niveau du gène ancestral G111. Ce gène fait partie des gènes présentant une co-apparition pour au moins deux fibulines humaines. Les PPI mises en jeu sont DYRK1A,

GFI1B, LTBP1, MEOX2, OTX1, TGFB1. Comme expliqué précédemment, parmi ces six protéines d'interaction, trois sont spécifiquement partagées par les 3 fibulines humaines mises en jeu.
Les interactions partagées par cinq fibulines ont permis une confirmation des prédictions réalisées par l'outil PastML. En effet, ces annotations remontent dans l'arbre jusqu'à retrouver l'ancêtre commun à ces 5 fibulines. À l'inverse, des annotations peu partagées, comme pour RECQL4 qui n'interagit qu'avec la fibuline-5 et l'hémicentine-2, ne remontent pas dans l'arbre.

4.3 Impact des outils et paramètre utilisés sur les prédictions

La variation des paramètres de Paloma a été réalisée dans les limites de ce que l'aide pour cet outil préconise. L'utilisation des paramètres par défaut permet de sélectionner des alignements locaux d'une longueur maximum de 20 ainsi qu'un seuil placé à 10 pour la filtration des alignements locaux, ce qui a conduit à l'identification de 508 modules. La diminution de la taille maximum des alignements locaux d'entrée (17 puis 15) montre une augmentation du nombre de modules identifiés.
La diminution du paramètre `t` pour une même longueur maximum montre une augmentation du nombre de modules identifiés. La diminution de ce paramètre permet une filtration des alignements vis-à-vis de leur similarité.
Pour tester l'impact des variations de ces paramètres, les gènes avec co-apparition ont été utilisés comme indicateur de robustesse. Au cours de l'étude, quatre nœuds ont été identifiés comme intéressants grâce à la présence de deux ou plus descendants humains. Ces gènes sont G50, G110, G111 et G115. Ces gènes d'intérêt ont été retrouvés à chaque variation de paramètre. En plus de ces quatre gènes, huit (G18, G35, G37, G46, G47, G68, G82, G94) ont été retrouvés à chaque variation de paramètre. À l'inverse, quatre gènes (G93, G96, G105 et G109) ne sont retrouvés que pour certains paramètres.
La présence des quatre gènes d'intérêt et des huit autres gènes montre une robustesse de ces résultats vis-à-vis de la variation de paramètres. Une observation plus précise des modules permettrait d'observer ceux présents malgré la variation des paramètres de paloma et leur position vis-à-vis des domaines identifiés par Pfam. Le choix des paramètres pour Paloma reste propre à la famille de protéines étudiées. L'utilisation d'une valeur basse pour `t` permet l'identification d'un plus grand nombre de modules. Cependant, certains modules peuvent ne pas être significativement intéressants, de par une similarité trop basse. L'augmentation de ce paramètre permettrait d'identifier moins de modules, mais ces derniers présenteraient une similarité plus forte que pour une valeur de `t` plus faible.

Pour réaliser les alignements de séquences multiples, c'est l'outil Muscle qui est implémenté dans le conteneur Docker de PhyloCharMod et qui est maintenant dans le conteneur Singularity que j'ai réalisé. Cependant, comme décrit par Olivier Dennler dans son article [18] et sa thèse [17], on peut utiliser l'outil PASTA. Il serait donc intéressant, pour notre étude sur les fibulines, de réaliser le calcul de l'arbre phylogénétique avec PASTA et de pouvoir ainsi le comparer avec celui obtenu avec Muscle. L'utilisation d'un outil tel que Muscle permet l'obtention de résultats plus rapidement, cependant, pour le cas de protéines longues, comme c'est le cas ici, l'utilisation d'un outil comme PASTA serait, en effet, plus adéquate.

5 Conclusion

Au cours de cette étude, j'ai réalisé un container Singularity permettant l'utilisation de l'outil PhyloCharMod, ainsi que plusieurs adaptations du code source. Cet outil a été développé afin d'identifier et de caractériser les motifs fonctionnels au sein de protéines multidomaines. Ce dernier a été réalisé sur les familles de protéines ADAMTS, mais n'a jamais été testé sur une autre famille de protéines. J'ai appliqué la version Singularity du pipeline d'analyse à la famille des fibulines, une famille de protéines impliquée dans la matrice extracellulaires dans le but de savoir si l'outil est généralisable à l'ensemble des protéines, mais aussi d'identifier les fonctions et motifs fonctionnels de cette famille. Cette étude a permis de calculer un arbre des gènes des fibulines chez huit espèces : *Homo sapiens, Mus musculus, Bos taurus, Gallus gallus, Xenopus tropicalis, Danio rerio, Ciona intestinalis, Caenorhabditis elegans*. Au cours de l'étude, il a été mis en évidence qu'aucune fibuline n'existait chez la *Drosophila melanogaster*. L'arbre a permis de montrer l'existence d'une duplication de la fibuline-7 au niveau de l'ancêtre des Chordata. L'une des fibulines issues de cette duplication a été perdue au niveau de l'ancêtre des Euteleostomi. Le réseau d'interaction des fibulines avec d'autres protéines a été réalisé et utilisé au sein de l'outil PhyloCharMod comme annotation fonctionnelle. L'histoire évolutive des modules conservés et des PPI a permis d'identifier des gènes ancestraux de co-apparition. Parmi ces derniers, quatre présentent deux ou plus descendants humains. Une variation des paramètres de l'outil de segmentation en module, Paloma, a permis de prouver la robustesse des résultats de co-apparition pour ces quatre gènes ancestraux.

References

[1] Alyssa Soles, Adem Selimovic, Kaelin Sbrocco, Ferris Ghannoum, Katherine Hamel, Emmanuel Labrada Moncada, Stephen Gilliat, and Marija Cvetanovic. "Extracellular Matrix Regulation in Physiology and in Brain Disease". In: *Int J Mol Sci* 24.8 (Apr. 2023), p. 7049. ISSN: 1422-0067. DOI: 10.3390/ijms24087049. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10138624/.

[2] Deviyani Mahajan, Sudhakar Kancharla, Prachetha Kolli, Amarish Kumar Sharma, Sanjeev Singh, Sudarshan Kumar, Ashok Kumar Mohanty, and Manoj Kumar Jena. "Role of Fibulins in Embryonic Stage Development and Their Involvement in Various Diseases". In: *Biomolecules* 11.5 (May 2021), p. 685. ISSN: 2218-273X. DOI: 10.3390/biom11050685. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8147605/.

[3] Clemens Rauer, Neeladri Sen, Vaishali P. Waman, Mahnaz Abbasian, and Christine A. Orengo. "Computational approaches to predict protein functional families and functional sites". en. In: *Current Opinion in Structural Biology* 70 (Oct. 2021), pp. 108–122. ISSN: 0959440X. DOI: 10.1016/j.sbi.2021.05.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959440X21000816.

[4] Constance J. Jeffery. "Current successes and remaining challenges in protein function prediction". en. In: *Front. Bioinform.* 3 (July 2023), p. 1222182. ISSN: 2673-7647. DOI: 10.3389/fbinf.2023.1222182. URL: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1222182/full.

[5] Yan Wang, Hang Zhang, Haolin Zhong, and Zhidong Xue. "Protein domain identification methods and online resources". en. In: *Computational and Structural Biotechnology Journal* 19 (2021). Publisher: Research Network of Computational and Structural Biotechnology, p. 1145. DOI: 10.1016/j.csbj.2021.01.041. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7895673/.

[6] Andrew Harrison, Frances Pearl, Ian Sillitoe, Tim Slidel, Richard Mott, Janet Thornton, and Christine Orengo. "Recognizing the fold of a protein structure". eng. In: *Bioinformatics* 19.14 (Sept. 2003), pp. 1748–1759. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg240.

[7] John A Capra, Roman A Laskowski, Janet M Thornton, Thomas A Funkhouser, and Mona Singh. "Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure — Supplementary Material". en. In: (2021).

[8] Dariya S. Glazer, Randall J. Radmer, and Russ B. Altman. "Improving structure-based function prediction using molecular dynamics". eng. In: *Structure* 17.7 (July 2009), pp. 919–929. ISSN: 1878-4186. DOI: 10.1016/j.str.2009.05.010.

[9] Sc Pakhrin, Shrestha B, Adhikari B, and Kc Db. "Deep Learning-Based Advances in Protein Structure Prediction". en. In: *International journal of molecular sciences* 22.11 (May 2021). Publisher: Int J Mol Sci. ISSN: 1422-0067. DOI: 10.3390/ijms22115553. URL: https://pubmed.ncbi.nlm.nih.gov/34074028/.

[10] J Jumper, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl Saa, Ballard Aj, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior Aw, Kavukcuoglu K, Kohli P, and Hassabis D. "Highly accurate protein structure prediction with AlphaFold". en. In: *Nature* 596.7873 (Aug. 2021). Publisher: Nature. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://pubmed.ncbi.nlm.nih.gov/34265844/.

[11] Christian J. A. Sigrist, Lorenzo Cerutti, Nicolas Hulo, Alexandre Gattiker, Laurent Falquet, Marco Pagni, Amos Bairoch, and Philipp Bucher. "PROSITE: A documented database using patterns and profiles as motif descriptors". In: *Briefings in Bioinformatics* 3.3 (Sept. 2002), pp. 265–274. ISSN: 1467-5463. DOI: 10.1093/bib/3.3.265. URL: https://doi.org/10.1093/bib/3.3.265.

[12] Rahel Schnellmann, Ragna Sack, Daniel Hess, Douglas S. Annis, Deane F. Mosher, Suneel S. Apte, and Ruth Chiquet-Ehrismann. "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". eng. In: *Mol Cell Proteomics* 17.7 (July 2018), pp. 1410–1425. ISSN: 1535-9484. DOI: 10.1074/mcp.RA118.000676.

[13] M Gribskov, A D McLachlan, and D Eisenberg. "Profile analysis: detection of distantly related proteins." en. In: *Proc. Natl. Acad. Sci. U.S.A.* 84.13 (July 1987), pp. 4355–4358. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.84.13.4355. URL: https://pnas.org/doi/full/10.1073/pnas.84.13.4355.

[14] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, Robert D. Finn, and Alex Bateman. "Pfam: The protein families database in 2021". eng. In: *Nucleic Acids Res* 49.D1 (Jan. 2021), pp. D412–D419. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa913.

[15] Goulven Kerbellec. "Learning automata modelling families of protein sequences". In: (June 2008).

[16] François Coste. *Protomata learner*. 2022. URL: https://tools.genouest.org/tools/protomata/learn/.

[17] Olivier Dennler. "Caractérisation en modules fonctionnels des protéines ADAMTS-TSL par approches de phylogénies". fr. PhD thesis. Université de Rennes, Dec. 2022. URL: https://theses.hal.science/tel-04051073.

[18] Olivier Dennler, François Coste, Samuel Blanquart, Catherine Belleannée, and Nathalie Théret. "Phylogenetic inference of the emergence of sequence modules and protein-protein interactions in the ADAMTS-TSL family". eng. In: *PLoS Comput Biol* 19.8 (Aug. 2023), e1011404. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011404.

[19] Jonathan A. Eisen. "Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis". en. In: *Genome Res.* 8.3 (Jan. 1998). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 163–167. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.8.3.163. URL: https://genome.cshlp.org/content/8/3/163.

[20] Duncan Brown and Kimmen Sjölander. "Functional Classification Using Phylogenomic Inference". en. In: *PLOS Computational Biology* 2.6 (June 2006). Publisher: Public Library of Science, e77. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.0020077. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0020077.

[21] Maureen Stolzer, Katherine Siewert, Han Lai, Minli Xu, and Dannie Durand. "Event inference in multidomain families with phylogenetic reconciliation". In: *BMC Bioinformatics* 16.14 (Oct. 2015), S8. ISSN: 1471-2105. DOI: 10.1186/1471-2105-16-S14-S8. URL: https://doi.org/10.1186/1471-2105-16-S14-S8.

[22] Lei Li and Mukul S. Bansal. "An Integrated Reconciliation Framework for Domain, Gene, and Species Level Evolution". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.1 (Jan. 2019). Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 63–76. ISSN: 1557-9964. DOI: 10.1109/TCBB.2018.2846253. URL: https://ieeexplore.ieee.org/document/8382181.

[23] Lei Li and Mukul S. Bansal. "Simultaneous Multi-Domain-Multi-Gene Reconciliation Under the Domain-Gene-Species Reconciliation Model". en. In: *Bioinformatics Research and Applications*. Ed. by Zhipeng Cai, Pavel Skums, and Min Li. Vol. 11490. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 73–86. ISBN: 978-3-030-20241-5 978-3-030-20242-2. DOI: 10.1007/978-3-030-20242-2_7. URL: https://link.springer.com/10.1007/978-3-030-20242-2_7.

[24] Robert C. Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". eng. In: *Nucleic Acids Res* 32.5 (2004), pp. 1792–1797. ISSN: 1362-4962. DOI: 10.1093/nar/gkh340.

[25] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0". en. In: *Systematic Biology* 59.3 (Mar. 2010), pp. 307–321. ISSN: 1076-836X, 1063-5157. DOI: 10.1093/sysbio/syq010. URL: https://academic.oup.com/sysbio/article/59/3/307/1702850.

[26] Yi-Chieh Wu, Matthew D. Rasmussen, Mukul S. Bansal, and Manolis Kellis. "TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees". en. In: *Systematic Biology* 62.1 (Jan. 2013), pp. 110–120. ISSN: 1076-836X, 1063-5157. DOI: 10.1093/sysbio/sys076. URL: https://academic.oup.com/sysbio/article/62/1/110/1657799.

[27] Sohta A Ishikawa, Anna Zhukova, Wataru Iwasaki, and Olivier Gascuel. "A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios". In: *Molecular Biology and Evolution* 36.9 (Sept. 2019), pp. 2069–2085. ISSN: 0737-4038. DOI: 10.1093/molbev/msz131. URL: https://doi.org/10.1093/molbev/msz131.

[28] Dirk Merkel. "Docker: lightweight linux containers for consistent development and deployment". In: *Linux journal* 2014.239 (2014), p. 2.

[29] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. "Singularity: Scientific containers for mobility of compute". en. In: *PLOS ONE* 12.5 (May 2017). Publisher: Public Library of Science, e0177459. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0177459. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177459.

[30] David M. Emms and Steven Kelly. "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy". In: *Genome Biology* 16.1 (Aug. 2015), p. 157. ISSN: 1474-760X. DOI: 10.1186/s13059-015-0721-2. URL: https://doi.org/10.1186/s13059-015-0721-2.

[31] David M. Emms and 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.

[32] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J. Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View". eng. In: *Methods Mol Biol* 1374 (2016), pp. 23–54. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-3167-5_2.

[33] Salvador Capella-Gutiérrez, José M. Silla-Martínez, and Toni Gabaldón. "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses". In: *Bioinformatics* 25.15 (Aug. 2009), pp. 1972–1973. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp348. URL: https://doi.org/10.1093/bioinformatics/btp348.

[34] Yannis Nevers, Arnaud Kress, Audrey Defosset, Raymond Ripp, Benjamin Linard, Julie D Thompson, Olivier Poch, and Odile Lecompte. "OrthoInspector 3.0: open portal for comparative genomics". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D411–D418. ISSN: 0305-1048. DOI: 10.1093/nar/gky1068. URL: https://doi.org/10.1093/nar/gky1068.

[35] Frédéric Lemoine, Damien Correia, Vincent Lefort, Olivia Doppelt-Azeroual, Fabien Mareuil, Sarah Cohen-Boulakia, and Olivier Gascuel. "NGPhylogeny.fr: new generation phylogenetic services for non-specialists". en. In: *Nucleic Acids Research* 47.W1 (July 2019), W260–W265. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz303. URL: https://academic.oup.com/nar/article/47/W1/W260/5480904.

[36] Ivica Letunic and Peer Bork. "Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool". In: *Nucleic Acids Research* (Apr. 2024), gkae268. ISSN: 0305-1048. DOI: 10.1093/nar/gkae268. URL: https://doi.org/10.1093/nar/gkae268.

[37] Bruno Aranda, Hagen Blankenburg, Samuel Kerrien, Fiona S L Brinkman, Arnaud Ceol, Emilie Chautard, Jose M Dana, Javier De Las Rivas, Marine Dumousseau, Eugenia Galeota, Anna Gaulton, Johannes Goll, Robert E W Hancock, Ruth Isserlin, Rafael C Jimenez, Jules Kerssemakers, Jyoti Khadake, David J Lynn, Magali Michaut, Gavin O'Kelly, Keiichiro Ono, Sandra Orchard, Carlos Prieto, Sabry Razick, Olga Rigina, Lukasz Salwinski, Milan Simonovic, Sameer Velankar, Andrew Winter, Guanming Wu, Gary D Bader, Gianni Cesareni, Ian M Donaldson, David Eisenberg, Gerard J Kleywegt, John Overington, Sylvie Ricard-Blum, Mike Tyers, Mario Albrecht, and Henning Hermjakob. "PSICQUIC and PSISCORE: accessing and scoring molecular interactions". In: *Nat Methods* 8.7 (June 2011), pp. 528–529. ISSN: 1548-7091. DOI: 10.1038/nmeth.1637. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3246345/.

[38] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". eng. In: *Genome Res* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303.

[39] Fernando Segade. "Molecular evolution of the fibulins: Implications on the functionality of the elastic fibulins". In: *Gene* 464.1 (Sept. 2010), pp. 17–31. ISSN: 0378-1119. DOI: 10.1016/j.gene.2010.05.003. URL: https://www.sciencedirect.com/science/article/pii/S0378111910002325.

[40] John S. Taylor, Ingo Braasch, Tancred Frickey, Axel Meyer, and Yves Van de Peer. "Genome duplication, a trait shared by 22000 species of ray-finned fish". eng. In: *Genome Res* 13.3 (Mar. 2003), pp. 382–390. ISSN: 1088-9051. DOI: 10.1101/gr.640303.

[41] Susana de Vega, Tsutomu Iwamoto, Takashi Nakamura, Kentaro Hozumi, Dianalee A. McKnight, Larry W. Fisher, Satoshi Fukumoto, and Yoshihiko Yamada. "TM14 is a new member of the fibulin family (fibulin-7) that interacts with extracellular matrix molecules and is active for cell binding". eng. In: *J Biol Chem* 282.42 (Oct. 2007), pp. 30878–30888. ISSN: 0021-9258. DOI: 10.1074/jbc.M705847200.

[42] Papiya Chakraborty, Shiba Prasad Dash, and Pranita P. Sarangi. "The role of adhesion protein Fibulin7 in development and diseases". In: *Mol Med* 26 (May 2020), p. 47. ISSN: 1076-1551. DOI: 10.1186/s10020-020-00169-z. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7238533/.

[43] Bruce E. Vogel, Joaquin M. Muriel, Chun Dong, and Xuehong Xu. "Hemicentins: what have we learned from worms?" eng. In: *Cell Res* 16.11 (Nov. 2006), pp. 872–878. ISSN: 1748-7838. DOI: 10.1038/sj.cr.7310100.

[44] Renato V. Iozzo and Liliana Schaefer. "Proteoglycan form and function: A comprehensive nomenclature of proteoglycans". eng. In: *Matrix Biol* 42 (Mar. 2015), pp. 11–55. ISSN: 1569-1802. DOI: 10.1016/j.matbio.2015.02.003.

[45] Gaetano Santulli. "Safety in numbers: Identifying multiple targets for beta cell proliferation". eng. In: *Sci Transl Med* 11.475 (Jan. 2019), eaaw5312. ISSN: 1946-6242. DOI: 10.1126/scitranslmed.aaw5312.

[46] Yang-Ling Li, Man-Man Zhang, Lin-Wen Wu, Ye-Han Liu, Zuo-Yan Zhang, Ling-Hui Zeng, Neng-Ming Lin, and Chong Zhang. "DYRK1A reinforces epithelial-mesenchymal transition and metastasis of hepatocellular carcinoma via cooperatively activating STAT3 and SMAD". eng. In: *J Biomed Sci* 29.1 (June 2022), p. 34. ISSN: 1423-0127. DOI: 10.1186/s12929-022-00817-y.

[47] Kimberly A. Brown, Amy-Joan L. Ham, Cara N. Clark, Nahum Meller, Brian K. Law, Anna Chytil, Nikki Cheng, Jennifer A. Pietenpol, and Harold L. Moses. "Identification of novel Smad2 and Smad3 associated proteins in response to TGF-beta1". eng. In: *J Cell Biochem* 105.2 (Oct. 2008), pp. 596–611. ISSN: 1097-4644. DOI: 10.1002/jcb.21860.

6 Annexes

6.1 Annexe 1

Figure 12. Arbre des espèces actuelles ainsi que leurs ancêtres. Les branches avec une taille plus importante correspondent aux branches ayant pour descendants actuels une des espèces étudiées, elles mêmes représentées en gras. Les branches bleues indiquent les descendants de Chordata. Les ancêtres de la lignée humaine sont indiqués au niveau des noeuds internes de l'arbre en rouge (d'après [17])

6.2 Annexe 2

Table 4. Tableau représentant les groupes de fibuline partageant des mêmes PPI

Fibuline	**PPI**
FBLN1-FBLN3-FBLN4-FBLN5	GFI1B MEOX2
FBLN1-FBLN2-FBLN3-FBLN4-FBLN5	VTN EMILIN1 MFAP4 MFAP5 NUFIP2 ATN1 HOXA1 FBN2 TFAP2C MFAP1 MFAP2 FBN1 ELN MFAP3
FBLN3-FBLN4	IGFBP6 BAG6 CLPP PROP1 LTBP3 ZNF768 SGTA ANAPC11 CYSRT1 FAM110A SHANK3 NOS3 RERE LCE2C ATXN7
FBLN3-FBLN5	ZNF696 ZNF747 MAGEA3 ZBTB9 ZNF627 ZNF324 DNAJC10 EGFL7 ZNF136 ZNF460
FBLN3-FBLN4-FBLN5	LTBP1 TGFB1 DYRK1A
FBLN1-FBLN5	IGFL3 PRNP ST14 ZFP41 ZNF764 ITM2B
FBLN1-FBLN4	MRPL12 CXCL5 CAVIN1 PLSCR1 PLSCR4 LCE1A LINGO1 KRTAP11-1 FXR1 TRIM42 LTBP4 ADAMTSL4
FBLN1-FBLN3-FBLN4	CACNA1A LCE1C LCE3A
FBLN1-FBLN4-FBLN5	CREB5 FBLN5
FBLN2-FBLN4	FBXW5 COL8A1
FBLN1-FBLN2	NID1 HSPG2 VCAN ACAN FN1
FBLN1-FBLN3	KRTAP19-2 ECM1 TIMP3 HTRA1 OCLN DDIAS NLGN3
FBLN2-FBLN3-FBLN4-FBLN5	OTX1
FBLN1-HMCN1	ARMS2
FBLN2-FBLN3-FBLN4	LCE3D HSD3B7
FBLN4-FBLN5	LOXL1 EFEMP2 LOX
FBLN1-FBLN3-FBLN5	FBLN1 ZNF444
FBLN5-FBLN7	ASPH
FBLN3	FBLN7 HSPA5
FBLN2-FBLN3-FBLN5	ZNF408
FBLN2-FBLN5	LAMA5
FBLN5-HMCN2	RECQL4

Elisa CHENEL Master 2 Bioinformatique – Université de Rennes (2023-2024)

Résumé

Les fibulines sont une famille de protéines multidomaines de la matrice extracellulaire, composée de huit membres. Une analyse protéomique par l'équipe DYMEC2 de l'IRSET a permis de mettre en évidence l'association entre trois fibulines et la gravité de la fibrose dans le cas de maladie chronique du foie. Leur fonction dans la pathologie reste peu connue. Un outil permettant la caractérisation des motifs fonctionnels par la détection de modules conservé via l'utilisation de l'inférence phylogénétique de l'histoire évolutive des espèces, des gènes, des modules et des fonctions chez neuf espèces représentant l'évolution des protéines a été implémenté par Olivier Dennler. Cet outil se base sur des méthodes d'homologie de séquence, de phylogénétique et de réconciliation. Cet outil a été utilisé pour réaliser une étude sur la famille des fibulines. Malgré ça, mise à disposition sous la forme d'un conteneur Docker, des modifications ont dû être réalisées pour une utilisation sur le cluster de calcul Genouest. L'étude a permis le calcul d'un arbre phylogénétique des fibulines et l'identification d'événements de duplication à l'origine des fibulines actuelles. Une de ces duplications a permis l'identification d'une fibuline absente chez les mammifères. La reconstruction de l'histoire évolutive de petites régions conservées (module) et de l'histoire évolutive des annotations a permis l'identification de quatre gènes ancestraux présentant une co-apparition module/annotation avec au moins quatre descendants humains. L'un de ces gènes présente une annotation indiquant une interaction avec le TGFB1, une protéine impliquée dans la fibrose.

Mots-clés : Fibuline, multidomaine, phylogénomique, réconciliation, évolution

Abstract

Fibulins are an eight-member family of multidomain extracellular matrix proteins. Proteomic analysis by IREST's DYMEC2 team has highlighted the association between three members of the fibulin family and the severity of fibrosis in chronic liver disease. Little is known about their function in the disease. To this end, Olivier Dennler has implemented a tool to characterize functional motifs by detecting conserved modules using phylogenetic inference of the evolutionary history of species, genes, modules and functions in nine species representing protein evolution. This tool is based on homology, sequence, phylogenetic and reconciliation methods. It was used to carry out a study on the fibulin family. However, since it was made available as a Docker container, modifications had to be made to use on the Genouest computing cluster. The study enabled the calculation of a fibulin phylogenetic tree and the identification of duplication events at the origin of today's fibulins. One of these duplications led to the identification of a fibulin absent in mammals. Reconstruction of the evolutionary history of small conserved regions (module) and of the evolutionary history of annotations enabled the identification of four ancestral genes showing module/annotation co-appearance with at least four human descendants. These genes remain present despite a variation in the parameters of one of the tool's dependencies.

Key words : Fibulin, multidomain, phylogenomics, reconciliation, evolution