

Projet Bio-Informatique et Modélisation

**Bases moléculaires du syndrome de
Marfan, focus sur la protéine géante
fibrilline 1**

Zaina Dali

Simon Liétar

sous la supervision de

Louis Carrel-Billiard

Elodie Laine

Mai 2024

Table des matières

Projet Bio-Informatique et Modélisation	1
Bases moléculaires du syndrome de Marfan, focus sur la protéine géante fibrilline 1	1
Introduction	3
Présentation de la fibrilline 1	3
Structure de la fibrilline 1	4
Mutations	5
Descripteurs structuraux	6
Surface accessible au solvant (SASA)	6
Variance circulaire	7
Flexibilité et écart à la structure consensus	9
DSSP	9
Descripteurs non-structuraux	10
GEMME	10
Score de polymorphisme	11
Interprétation	12
Corrélations entre descripteurs	12
Analyse en composantes principales	13
Classification en fonction de la pression de sélection et de l'exposition	16
Conclusion	17
Bibliographie	18

Introduction

Le syndrome de Marfan est une maladie génétique relativement rare présente chez l'Homme et qui affecte le tissu conjonctif. Elle est caractérisée par une variété de manifestations cliniques, notamment des anomalies cardiaques, ophtalmologiques et squelettiques. Bien que des progrès aient été réalisés dans la compréhension et la prise en charge de cette maladie, des défis persistent, avec notamment une morbidité substantielle et une mortalité prématuée qui demeure. Au niveau moléculaire, le syndrome de Marfan est principalement causé par des mutations dans le gène *FBN1* codant pour la protéine fibrilline 1, une composante essentielle de la matrice extracellulaire.

Ce rapport se concentre sur l'étude des bases moléculaires du syndrome de Marfan, en mettant l'accent sur la protéine géante fibrilline 1. Notre objectif est d'établir une classification structurale, évolutive et fonctionnelle des mutations faux sens observées dans la fibrilline 1. Cette classification revêt une importance capitale pour mieux comprendre la relation entre le génotype et le phénotype associés au syndrome de Marfan, ainsi que pour explorer les implications cliniques de ces mutations.

Le travail sur ce projet s'est fait en collaboration avec Pauline Arnaud, Nadine Hannah et Laurent Gouya, praticiens de l'hôpital Bichat qui travaillent sur la compréhension du syndrome de Marfan et sont à l'origine du projet. Ceux-ci nous ont fourni une liste de phénotypes associés à des mutations observées chez des patients.

La première phase du projet a consisté en l'extraction de descripteurs pouvant servir à la classification des mutations. La seconde phase du projet a eu pour but de comprendre les relations entre ces descripteurs pour arriver à une classification non supervisée, et de mettre celle-ci en parallèle avec les phénotypes connus.

En combinant des approches bio-informatiques, structurales et évolutives, ce rapport propose une démarche multidimensionnelle pour cartographier les mutations de la fibrilline 1.

Présentation de la fibrilline 1

La protéine fibrilline 1 est une protéine de 2871 résidus qui agit comme composante majeure de la matrice extracellulaire (ECM). Cette protéine est codée par le gène *FBN1* de 257 kbp localisé sur le chromosome 15 [1], [2].

Malgré sa taille importante, cette protéine est essentiellement composée de domaines semblables [3] :

- 4 domaines **EGF** semblables à l'epidermal growth factor qui contiennent chacun 6 cystéines.
- 43 domaines **EGFCB** (CB for calcium-binding) semblables à l'epidermal growth factor, mais capables de contenir un atome de calcium à leur extrémité N-terminale. Ces domaines contiennent également 6 cystéines chacun et sont très abondants dans les protéines structurales de l'ECM.
- 9 domaines **TB** pour TGF β -binding protein-like. Ces domaines contiennent chacun 8 cystéines.

Nous utilisons ici sur les annotations d'UniProt, mais la classification de ces domaines peut varier d'une base de données à l'autre. La répartition des domaines dans la protéine est donnée dans les figures suivantes dont la Fig. 1.

La fibrilline 1 a de nombreux partenaires d'interaction tels que les protéines MFAP, la fibrilline 2 ou encore la fibrilline 1 elle-même [4]. L'étude de ces interactions permettrait sûrement d'en savoir plus sur son fonctionnement, mais nous avons choisi de concentrer notre attention sur la protéine seule.

La présence de ces partenaires d'interactions laisse penser que les domaines semblables ont des rôles différents malgré leur apparence similaire. Nous avons donc essayé d'exploiter les comparaisons entre domaines pour comprendre le rôle de leurs différences.

Structure de la fibrilline 1

Il n'existe pas de structure expérimentale complète de la fibrilline 1, mais seulement des structures de sous-ensembles comprenant quelques domaines [2]. Nous avons utilisé AlphaFold 2 [5] puis 3 [6] pour obtenir une structure de la protéine. Bien que la mesure de confiance locale de la prédiction (pLDDT) soit correcte avec AlphaFold 3 (entre 70 et 90 sur 100 pour la majorité de la protéine, voir Fig. 1), l'erreur de la position par rapport aux autres résidus (PAE) est trop importante (30 Å, voir Fig. 2) pour pouvoir conclure sur la forme générale de la protéine et sur les interactions entre les domaines. En particulier, AlphaFold a tendance à produire une structure globulaire, là où la littérature donne plus de crédit à une structure linéaire et rigide [4].

Les régions hors des domaines EGF, EGFCB et TB ont toujours un pLDDT très faible (~ 20), probablement en raison de leur absence des protéines dont la structure expérimentale est connue.

En raison de la faible fiabilité des prédictions de la protéine entière, nous avons utilisé AlphaFold 2 pour prédire la structure des domaines individuellement, en incluant également les deux domaines adjacents afin d'avoir un « contexte » qui améliore la prédiction. En conséquence, les régions entre des domaines ne sont pas prédites ni analysées dans le reste du projet. Le pLDDT est meilleur que les prédictions de la protéine entière et le PAE assez bon au sein de chaque domaine (voir Fig. 3). Nous avons également testé ESMFold [7] mais obtenu des résultats moins bons qu'avec AlphaFold.

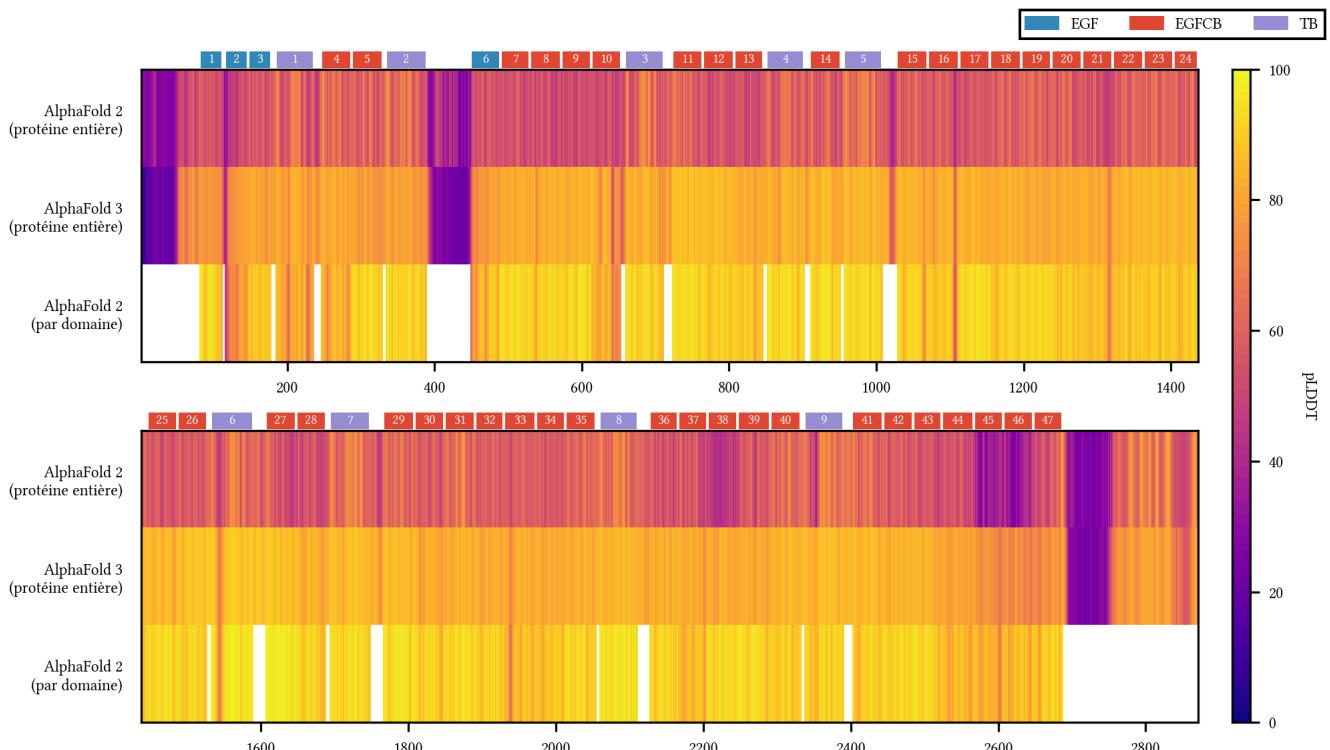


Fig. 1 pLDDT pour différentes prédictions

Nous avons inclus le pLDDT dans les descripteurs car celui-ci peut servir comme reflet de la structure secondaire de la protéine.

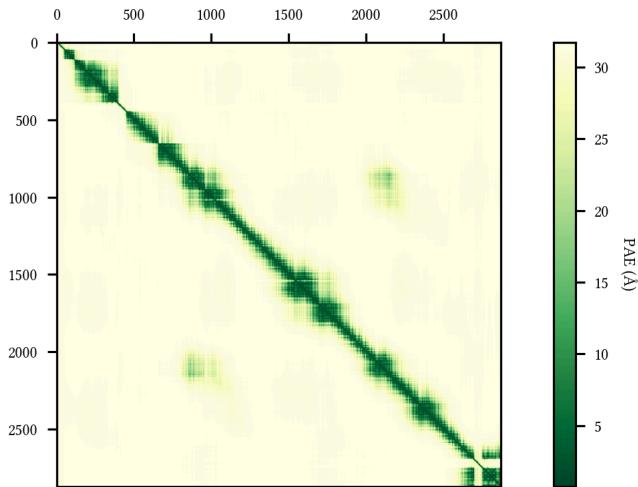


Fig. 2 Erreur d'alignement prédictée (PAE) sur la prédiction d'AlphaFold 3 pour la protéine complète

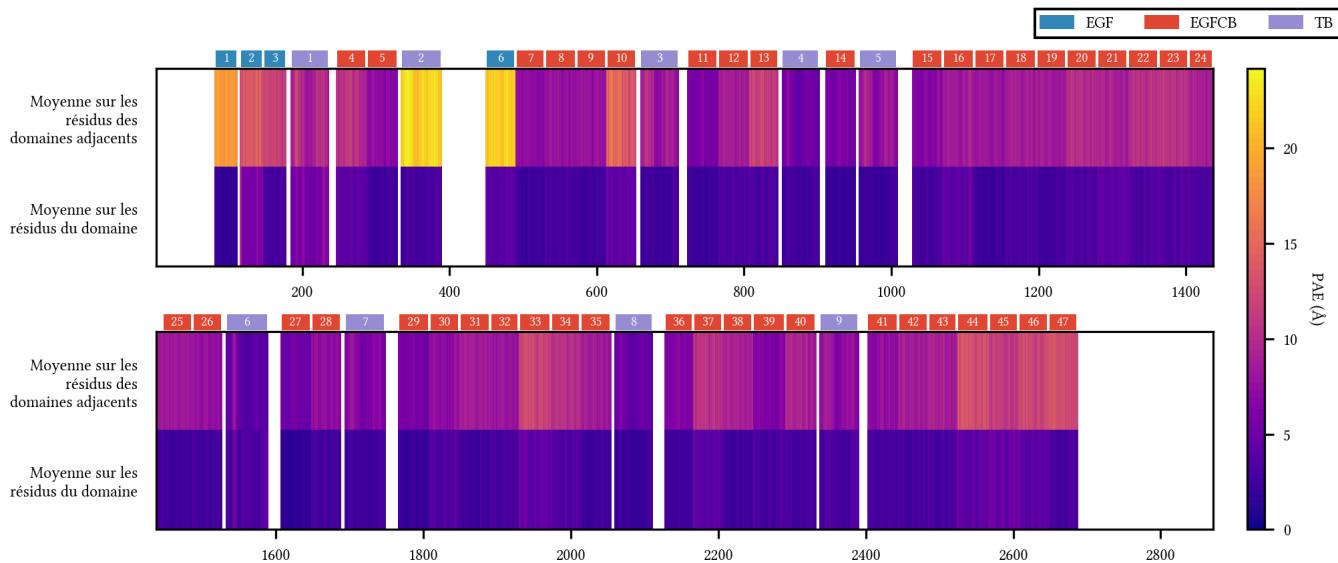


Fig. 3 Erreur d'alignement prédictée (PAE) sur les prédictions d'AlphaFold 2 pour les domaines individuels et les domaines adjacents

Mutations

De nombreuses mutations de *FBN1* sont connues comme responsables du syndrome de Marfan, et une mutation sur une seule des deux copies du gène est suffisante pour le provoquer. Le phénotype d'une mutation est toutefois très variable.

Les médecins de l'hôpital Bichat nous ont fourni une liste de 731 mutations faux sens observées sur des patients et comportant chacune zéro ou plus de 6 phénotypes, comme suit :

Effet	Nombre de mutations
Pneumothorax	8
Problème cardiaque*	24/486
Problème cutané	21
Problème ophtalmologique*	16/434
Problème neurologique	6
Problème squelettique*	26/485
Problème grave	46

Tableau 1 Effets recensés des mutations dans *FBN1*. Les effets marqués d'une astérisque (*) sont donnés avec deux niveaux de confiance ; la seconde valeur correspond au niveau le plus élevé.

Dans la suite du projet, on considérera comme pathogènes les mutations ayant au moins un effet avec le niveau de confiance le plus élevé, le cas échéant. Cela concerne 584 mutations sur 421 résidus.

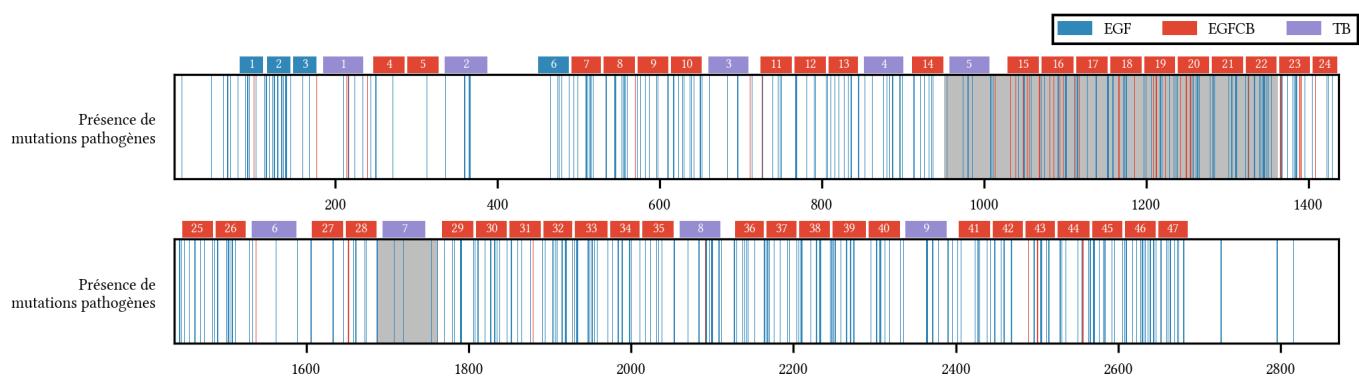


Fig. 4 Positions avec des mutations pathogènes. Les traits bleus représentent de mutations pathogènes et rouges pathogènes et graves. Les surfaces en gris sont les régions d'intérêt pour les médecins de l'hôpital Bichat.

Les mutations sont distribuées comme montré à la Fig. 4. Celle-ci est assez uniformément distribuée, sauf pour les régions entre les domaines où elles sont moins nombreuses.

Les médecins de l'hôpital Bichat ont déterminé deux régions comme particulièrement importantes. La première, dite région néonatale, contient de nombreuses mutations graves qui sont responsables de troubles dès la naissance. La seconde, dite TB 5 (TB 7 dans les annotations UniProt que nous utilisons), n'a pas de signature spécifique sur la Fig. 4, mais est caractéristique de symptômes particuliers du syndrome de Marfan, la dysplasie géléophysique ou acromicrique.

Descripteurs structuraux

Surface accessible au solvant (SASA)

Il est probable que les mutations pénètrent dans la surface de la protéine, donc responsable d'interactions, ou bien enfouies et donc affectant la structure interne de la protéine. Pour mesurer quantitativement cette propriété, nous avons utilisé la surface accessible au solvant (solvent-accessible surface area, SASA) [8].

Pour ce faire, nous avons employé la bibliothèque FreeSASA [9] qui implémente l'algorithme de Shrake–Rupley [10]. Celui-ci consiste à faire « rouler » une boule contre la surface de Van der Waals des atomes de la protéine. L'algorithme retourne, pour chaque résidu, un ratio entre la surface accessible au solvant dans cette structure et la surface théorique maximale qui est une fonction du type d'acide aminé.

Variance circulaire

Nous avons utilisé la variance circulaire comme une autre métrique pour caractériser l'emplacement d'un résidu dans la protéine. La variance circulaire de l'atome a est un nombre entre 0 et 1 défini comme :

$$CV_a = 1 - \frac{1}{N} \left\| \sum_{\substack{i \neq a \\ \|\vec{x}_i - \vec{x}_a\| < c}}^N \frac{\vec{x}_i - \vec{x}_a}{\|\vec{x}_i - \vec{x}_a\|} \right\|^2$$

où \vec{x}_i est la position de l'atome i .

La variance circulaire d'un résidu est définie comme la moyenne des variances circulaires de ses atomes.

Le seuil de distance c (en Å) permet de contrôler quels atomes sont considérés comme faisant partie du voisinage de l'atome a . Pour $c = \infty$, tous les atomes sont considérés comme faisant partie du voisinage et CV_a donne une mesure de l'enfouissement dans la protéine entière.

L'interprétation de la variance circulaire est la suivante: si celle-ci est à 1 comme en rouge dans la Fig. 5, alors tous les vecteurs $\vec{x}_i - \vec{x}_a$ se sont annulés et l'atome est donc au centre de son voisinage. Si celle-ci est à 0, tous les vecteurs vont dans la même direction est l'atome est à la surface de la protéine.

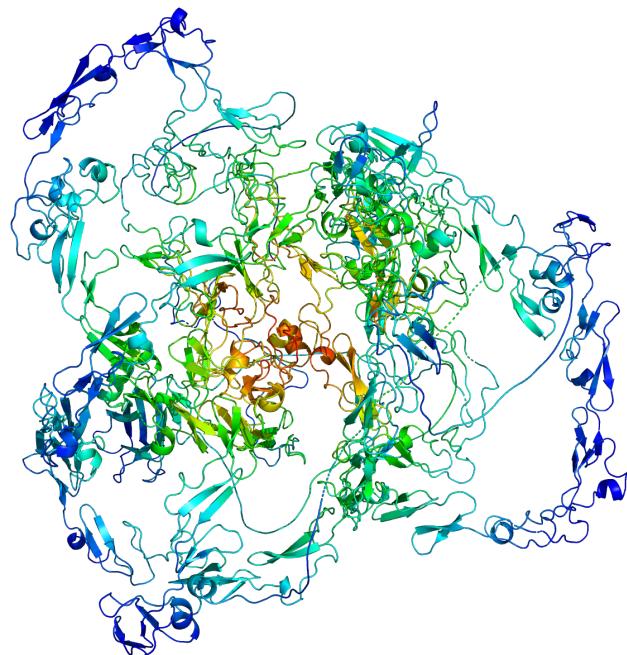


Fig. 5 Fibrilline 1 colorée avec la variance circulaire avec $c = \infty$. Rouge, au centre: 1.0, bleu foncé, en surface: 0.0. La structure ne sert que comme illustration car elle n'est pas fiable (voir Fig. 1).

L'interprétation exacte dépend du seuil choisi. Un seuil infini n'est pas souhaitable dans notre cas car la structure de la protéine entière n'est pas fiable. En revanche, des seuils à 10 et 20 Å permettent de capturer le contraste de l'enfouissement des résidus à l'échelle de la structure secondaire ou d'un domaine.

Pour implémenter cet algorithme efficacement sur une grande protéine, nous avons écrit un shader en WGL qui calcule la variance circulaire de tous les atomes en même temps sur le GPU. Nous avons publié [un paquet Python](#) et [un paquet Rust](#) pour utiliser l'algorithme sur n'importe quelle protéine en ligne de commande ou programmatiquement. D'autres algorithmes pourraient être utilisés, par exemple une grille qui permette de rapidement déterminer quels atomes sont dans le voisinage de l'atome considéré.

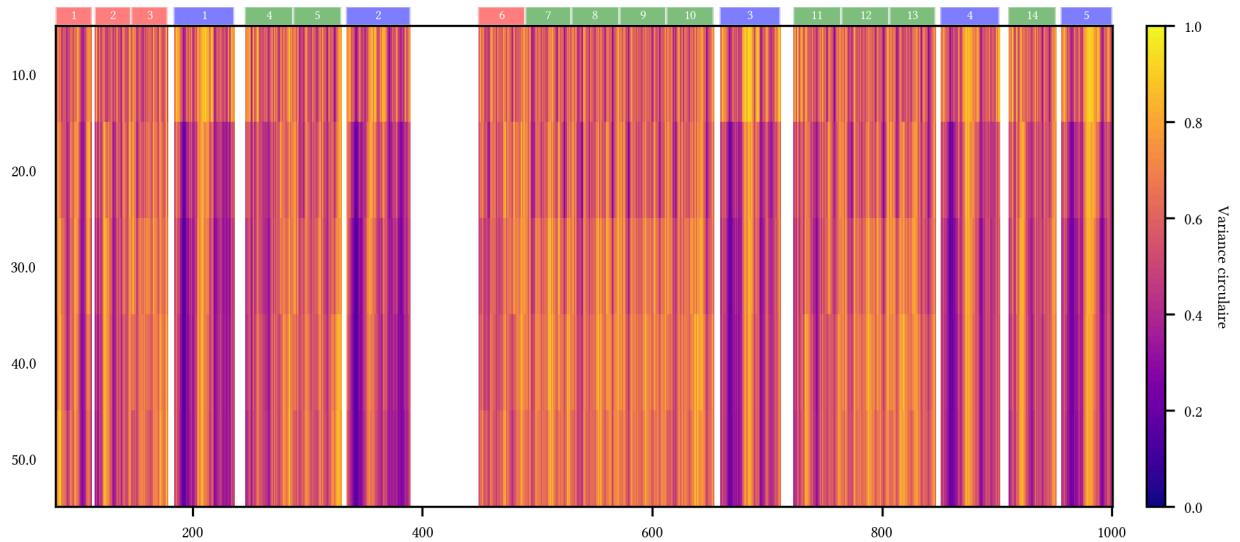


Fig. 6 Variance circulaire pour différents seuils de distance donnés sur l'axe vertical

On observe à la Fig. 6 que la variance circulaire tend à se stabiliser à mesure que de plus en plus d'atomes sont inclus, jusqu'à ce que le seuil de distance atteigne la plus grande distance entre deux atomes.

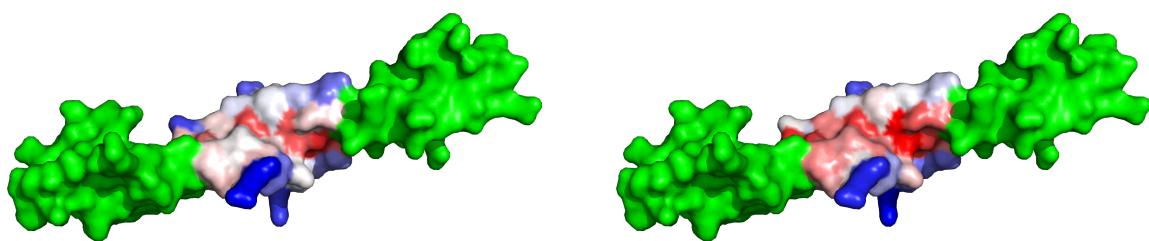


Fig. 7 Structure de 3 domaines EGFCB adjacents colorés avec la variance circulaire, avec des seuils de 10 Å (gauche) et 30 Å (droite). Les régions en vert sont les domaines adjacents qui sont considérés lors du calcul du domaine central mais dont la variance circulaire n'est pas calculée.

Flexibilité et écart à la structure consensus

La présence de nombreuses répétitions des mêmes domaines donne la possibilité de comparer leurs structures. En alignant les séquences des domaines d'un type donné, puis leurs structures, on peut estimer la position « consensus » \vec{x}_i de chaque résidu i pour un type de domaine k :

$$\vec{x}_i = \frac{1}{N} \sum_{d \in \mathcal{A}_k}^N \vec{x}_{d,i}$$

où $\vec{x}_{d,i}$ correspond à la position moyenne (ou alternativement à la position de l'atome de carbone α) des atomes du résidu i du domaine d , et \mathcal{A}_k à l'ensemble des domaines de type k .

On peut maintenant définir un nouveau descripteur pour chaque résidu comme l'écart à la position consensus correspondante. L'idée est que si l'un des domaines a obtenu une nouvelle fonction suite à mutation, celle-ci peut se refléter dans sa structure qui s'est alors éloignée de la position consensus.

On peut également calculer la distance moyenne de chaque domaine à la position consensus, pour chaque résidu i d'un type de domaine, ce qui donne la flexibilité F_i :

$$F_i = \frac{1}{N} \sum_{d \in \mathcal{A}_k}^N \|\vec{x}_{d,i} - \vec{x}_i\|$$

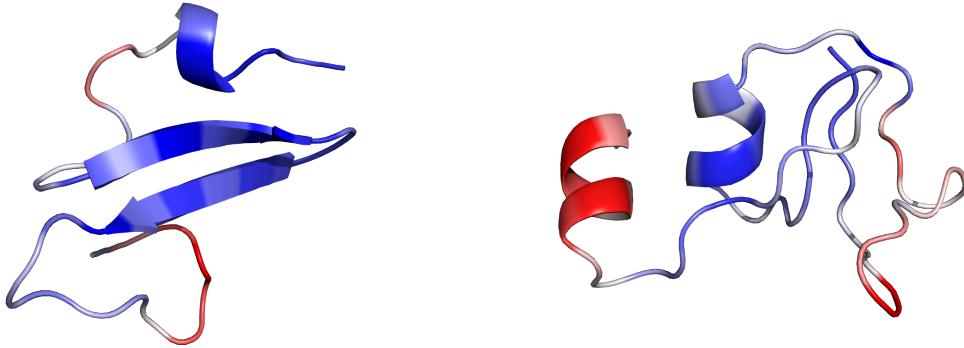


Fig. 8 Structures de domaines EGFCB (gauche) et TB (droite) colorés avec la flexibilité. Gauche : bleu foncé 0.37 Å, rouge 3.80 Å. Droite : bleu foncé 0.53 Å, rouge 7.26 Å.

On observe à la figure Fig. 8 que les hélices α et feuillets β sont les parties les moins flexibles des domaines, ce qui est attendu car ces structures secondaires sont plus sensibles aux changements de structure. Une exception notable est l'hélice α C-terminale des domaines TB qui est montrée comme flexible en raison de prédictions très éloignées dans quelques domaines.

DSSP

DSSP (Dictionary of Secondary Structure of Protein) [11] est un algorithme standard largement utilisé pour attribuer la structure secondaire aux acides aminés d'une protéine à partir de ses coordonnées à résolution atomique. Pour cela, DSSP utilise un dictionnaire de liaisons hydrogène et des caractéristiques géométriques pour attribuer la structure secondaire en analysant les liaisons hydrogène du squelette protéique ainsi que la topologie des feuillets β pour chaque résidu. Le

résultat obtenu est un code unique attribué à chaque résidu, indiquant sa structure secondaire, comme par exemple H pour une α -hélice, B pour un pont β isolé, E pour un brin étendu, et ainsi de suite. Nous avons appliqué l'algorithme DSSP [12] à la fibrilline 1 et avons obtenu les résultats présentés dans la figure suivante.

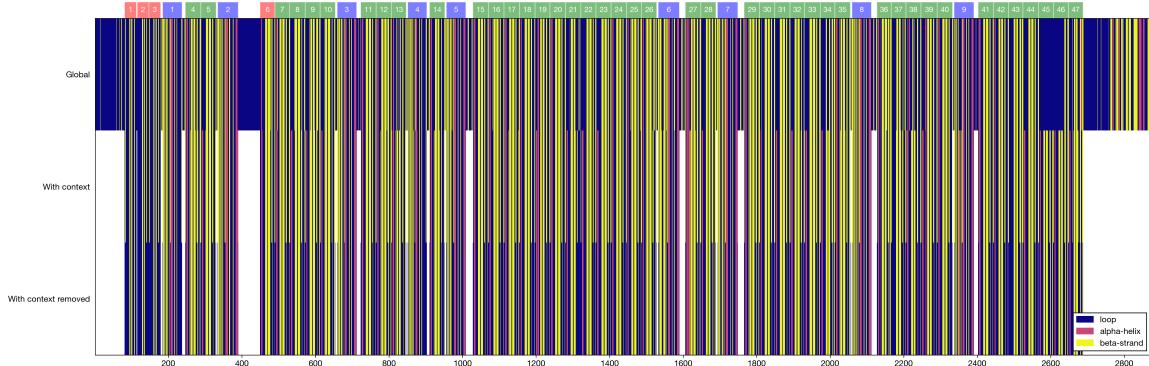


Fig. 9 Le pourcentage de résidus mutés présents dans chaque structure secondaire de la fibrilline 1

Comme le montre la figure, nous remarquons que la structure majoritaire de notre protéine est la structure en boucle (LOOP), qui contient 43,09 % des résidus mutés présents dans notre protéine. Ensuite, nous observons la structure en feuillet bête (BETA STRAND), qui est moins fréquente que la boucle, mais qui représente 47,2 % des résidus mutés. Enfin, la structure la moins présente est celle avec 9,71 % de résidus mutés.

Descripteurs non-structuraux

GEMME

Nous avons utilisé l'algorithme GEMME [13] pour obtenir un descripteur qui soit représentatif de la pression évolutive de chaque mutation. Pour ce faire, GEMME se fonde sur un alignement multi-séquence de la séquence d'intérêt et construit un modèle évolutif pour prédire l'effet de chaque acide aminé à chaque position.

Le résultat de GEMME étant donné sous forme d'une matrice avec le score pour chaque acide aminé pour chaque résidu, nous avons réduit cette matrice en deux descripteurs : le score GEMME moyen à chaque position et le score GEMME de la pire mutation à chaque position. Un score élevé indique une pression évolutive faible.

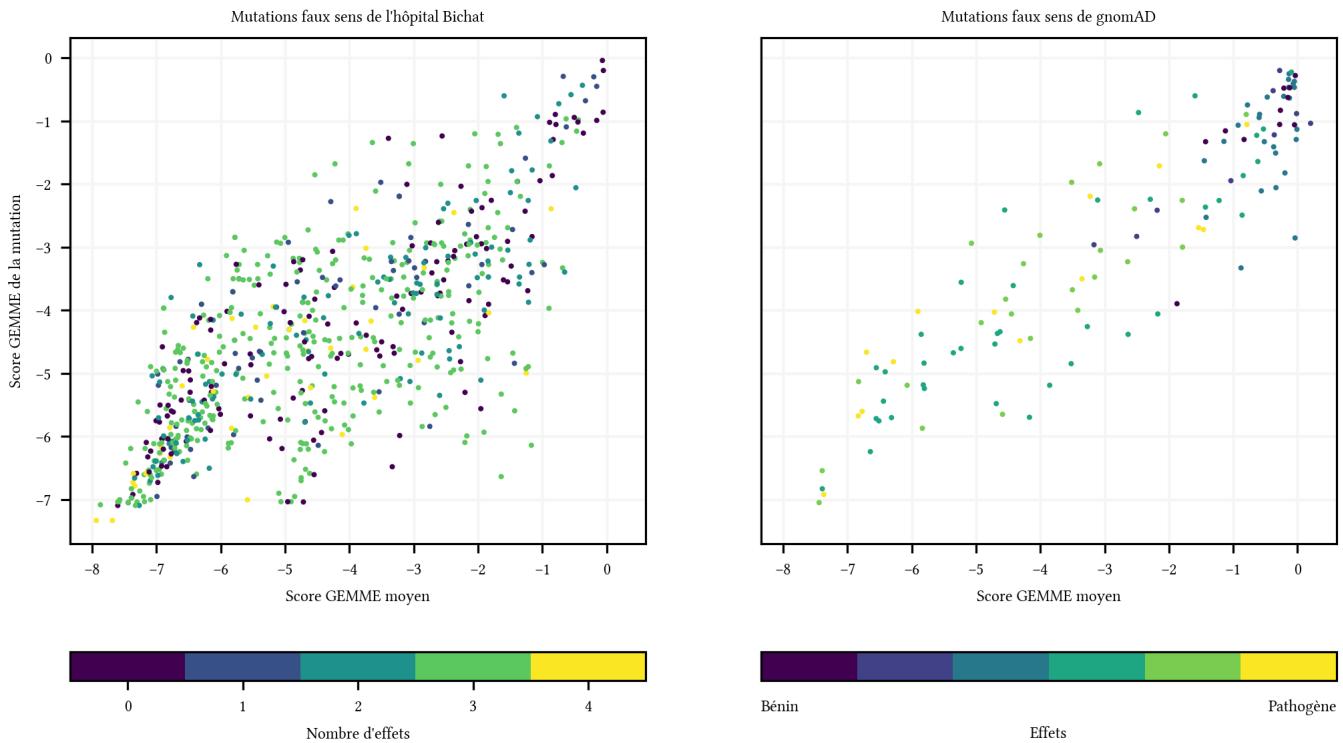


Fig. 10 Scores GEMME pour différentes mutations

Nous avons aussi calculé le score GEMME en ne considérant que les séquences annotées comme orthologues dans 156 espèces et donc assurément liées évolutivement à *FBN1*. En soustrayant les deux scores, on obtient une mesure différence entre les distances évolutives d'une mutation selon les deux contextes, que l'on utilisera comme descripteur noté Δ GEMME. Une valeur élevée correspond à une pression de sélection faible dans le contexte complet et élevée dans le contexte restreint limité aux orthologues.

La Fig. 10 montre que les mutations sur lesquelles nous nous basons ont généralement un score faible (< 2) voire très faible, quel que soit le nombre d'effets recensé. Les mutations recensées sur gnomAD, dont peu ont un phénotype connu, ont également tendance à avoir un score faible lorsque qu'elle sont annotées comme pathogènes.

L'outil PRESCOTT [14] se base sur GEMME et améliore les prédictions en prenant en compte les informations structurales provenant d'AlphaFold et les fréquences alléliques provenant de gnomAD [15]. Nous n'avons pas utilisé PRESCOTT en raison de la faible fiabilité des prédictions AlphaFold pour la fibrilline 1.

Score de polymorphisme

Afin de représenter la diversité des mutations connues, nous avons ajouté un « score de polymorphisme » qui est dérivé, pour chaque résidu, du nombre de mutations faux sens observées dans la base de données gnomAD [15]. Le but étant de mesurer quels résidus sont abondants sans être dangereux, nous n'avons pas compté les résidus annotés comme « pathogènes », « pathogènes/probablement pathogènes » ou « probablement pathogènes ».

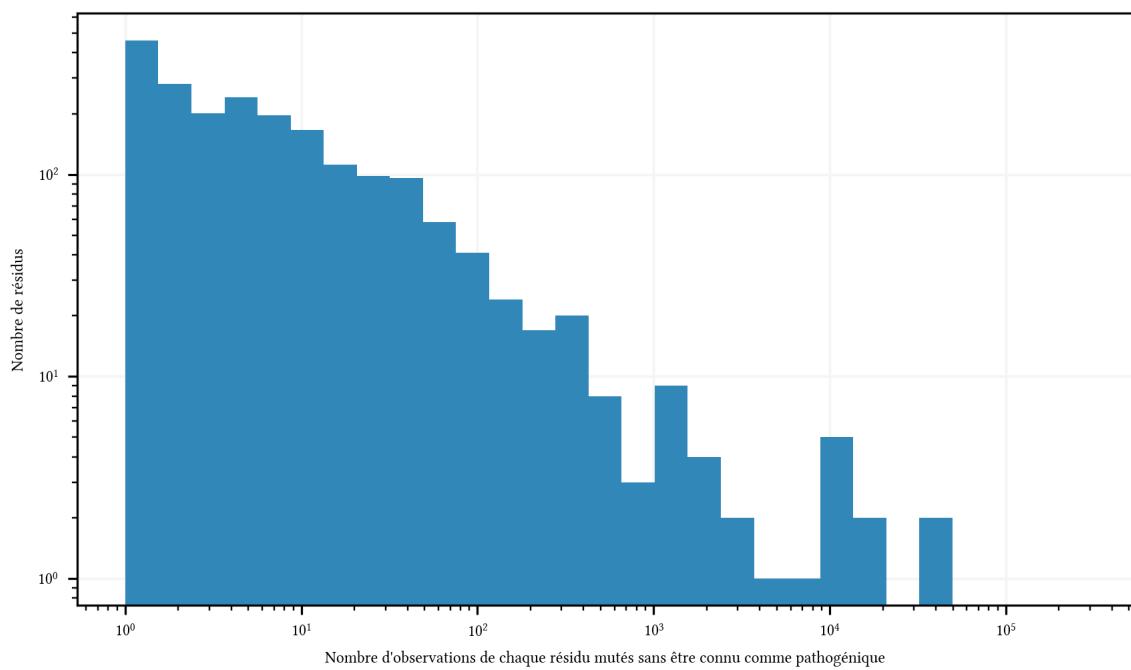


Fig. 11 Histogramme des observations

Nous avons utilisé une échelle logarithmique pour pallier aux fréquences d'apparitions distribuées inégalement.

Interprétation

Corrélations entre descripteurs

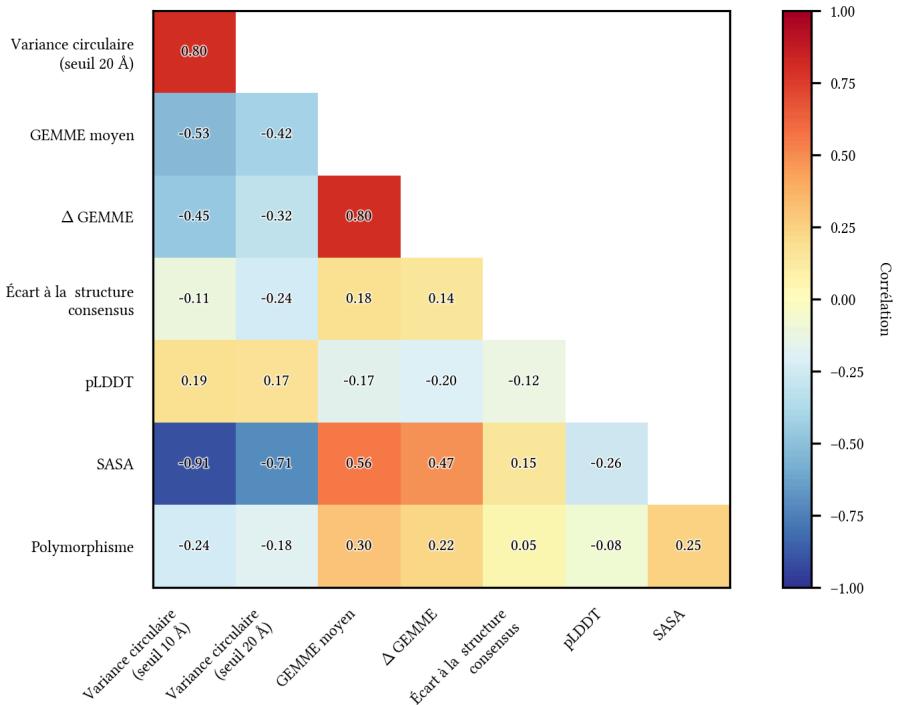


Fig. 12 Corrélations entre descripteurs

L'étude des corrélations à la Fig. 12 montre de fortes corrélations entre les descripteurs les variances circulaires à 10 et 20 Å (+0.80), ainsi que le score GEMME et le descripteur Δ GEMME calculé à partir des scores GEMME (+0.80). On observe également une corrélation importante entre la variance circulaire avec un faible seuil et la surface accessible au solvant (-0.91). Ceci est attendu car ces deux variables mesurent des propriétés semblables des résidus bien qu'avec des méthodes différentes. La corrélation est négative car la SASA est élevée et la variance circulaire faible lorsqu'un résidu est en surface.

Une autre corrélation notable est la correspondance entre le score GEMME moyen et la SASA (+0.56). Ceci révèle que les résidus en surface ont tendance à avoir une faible pression de sélection. Enfin, on note une corrélation entre le score de polymorphisme et le score GEMME moyen (+0.30), qui s'explique facilement par la faible pression de sélection, donc score GEMME élevé, dans les positions polymorphiques.

Analyse en composantes principales

Nous avons effectué des analyses en composantes principales (PCA) pour réduire la dimensionnalité des données et en obtenir une visualisation. Plus précisément, nous avons effectué une PCA par type de domaine pour prendre en compte les différences possibles entre eux. La PCA ne prend en compte que les positions pathogènes, mais toutes les positions sont affichées dans la Fig. 13. Les données sont mises à l'échelle avant la PCA. Nous n'avons pas inclus de variables discrètes, tel que la structure secondaire, car celles-ci donnent des résultats peu interprétables dans une PCA.

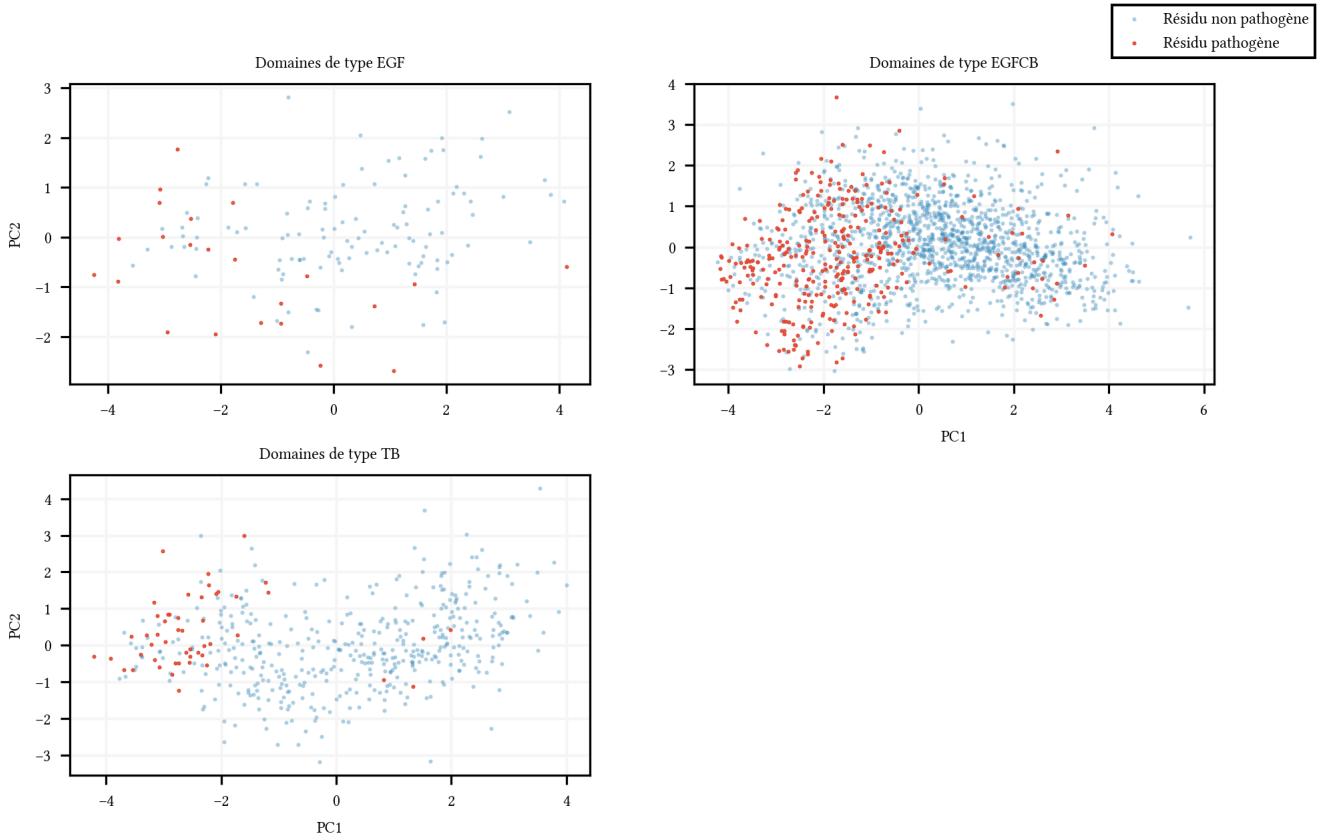


Fig. 13 Composantes PC1 et PC2 de PCA par type de domaine. Les points rouges correspondent à des positions pathogènes.

Il n'y pas de clusters évidents mais on observe une nette tendance pour les positions pathogènes à se trouver dans un sous-ensemble de l'espace.

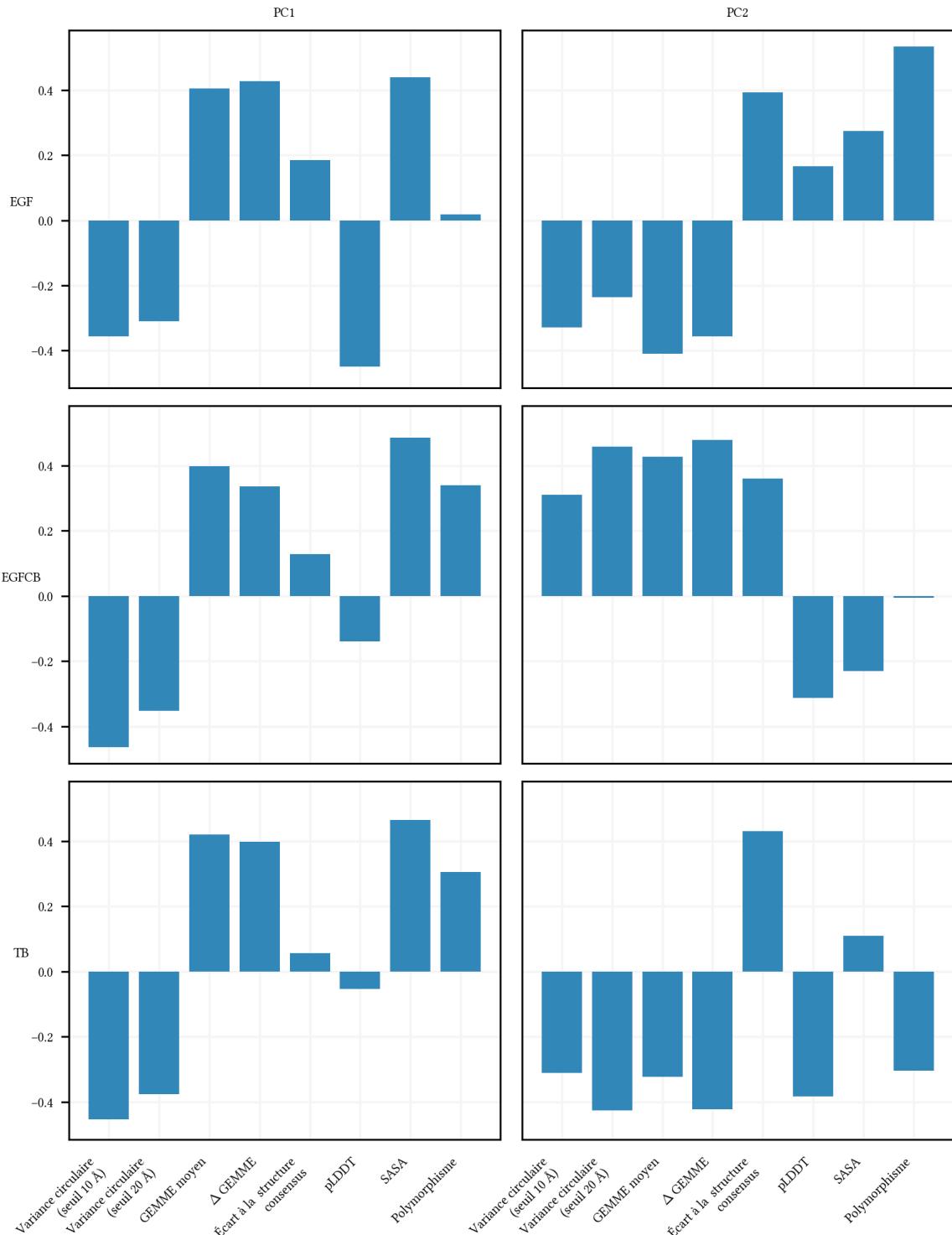


Fig. 14 Contribution des descripteurs aux composantes PC1 et PC2

Les contributions des descripteurs sont semblables d'un type de domaine à l'autre, mis à part quelques différences. Le premier mode (PC1), avec ~ 45 % de variance expliquée, donne un poids important à la pression de sélection (GEMME, polymorphisme) et à l'exposition (SASA, variance circulaire), et regroupe les descripteurs corrélés, tels que les variances circulaires.

Une valeur élevée de PC1 correspond à une position en surface avec une faible pression de sélection. La plupart des positions ayant une valeur faible de PC1 (et plus faible que la moyenne des positions),

on conclut que celles-ci sont généralement enfouies et avec une forte pression de sélection. Les domaines EGF ont également une forte contribution du pLDDT dont la signification n'est pas facilement interprétable.

Le mode PC2 explique ~ 22 % de la variance. Une valeur élevée de PC2 dans les domaines EGF et TB, ou faible dans les domaines EGFCB, correspond à une position en surface mais avec une pression de sélection élevée. Ce mode donne aussi une contribution non négligeable à l'écart avec la structure consensus selon les types de domaines. Il est toutefois difficile de conclure sur une classification à partir de cette composante.

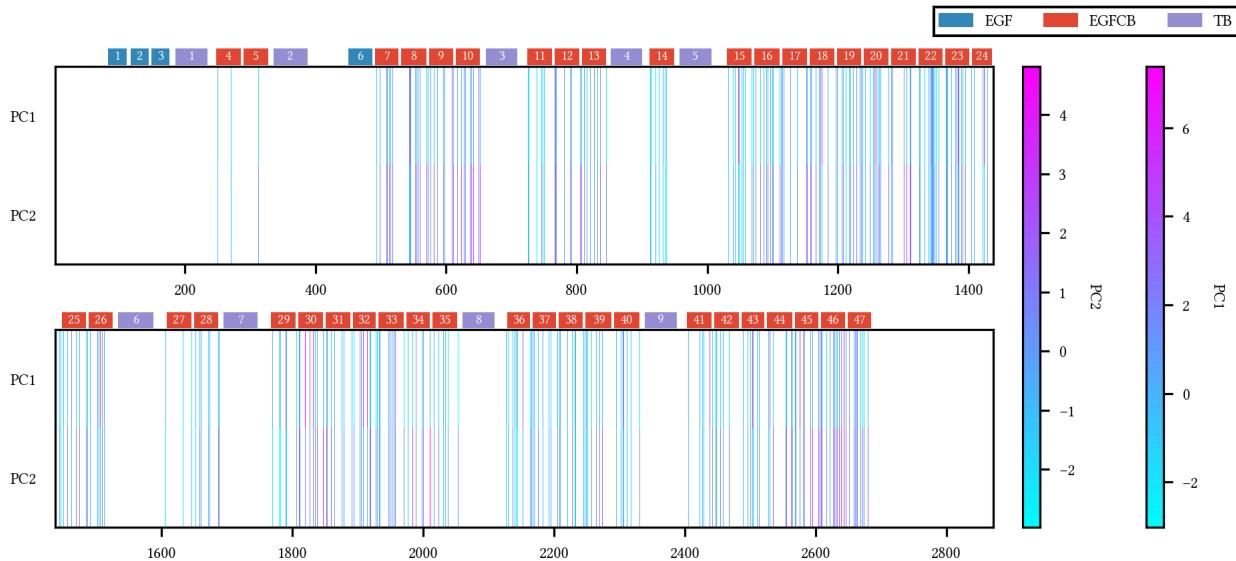


Fig. 15 Composantes PC1 et PC2 des mutations des domaines EGFCB, localisées dans la protéine

L'étude de la distribution des valeurs PC1 et PC2 dans les positions pathogènes des domaines EGFCB de la protéine (Fig. 15) montre que PC1 est généralement faible sauf pour quelques valeurs extrêmes, dans les domaines 30 et 32 par exemple. Pour PC2, les valeurs sont plus uniformes au sein d'un domaine, avec des domaines avec une valeur souvent faible de PC2 (par exemple 19 et 29), et d'autres une valeur élevée (par exemple 9, 10 et 46). Une analyse plus poussée sur les positions des mutations et la structure correspondante serait nécessaire pour conclure.

Classification en fonction de la pression de sélection et de l'exposition

Nous proposons une classification plus simple, sur la base de la pression de sélection (score GEMME moyen) et de l'exposition (variance circulaire), avec les classes suivantes :

- résidus enfouis (variance circulaire > 0.5) avec une forte pression de sélection (GEMME < -2.5), classe qui comprend la grande majorité des mutations (en bleu dans la Fig. 16) ;
- résidus enfouis avec une faible pression de sélection (en violet) ;
- résidus en surface avec une faible pression de sélection (en jaune).

Les deux dernières classes contiennent chacune une vingtaine de mutations, et sont intéressantes parce qu'elles contrastent avec le reste des positions.

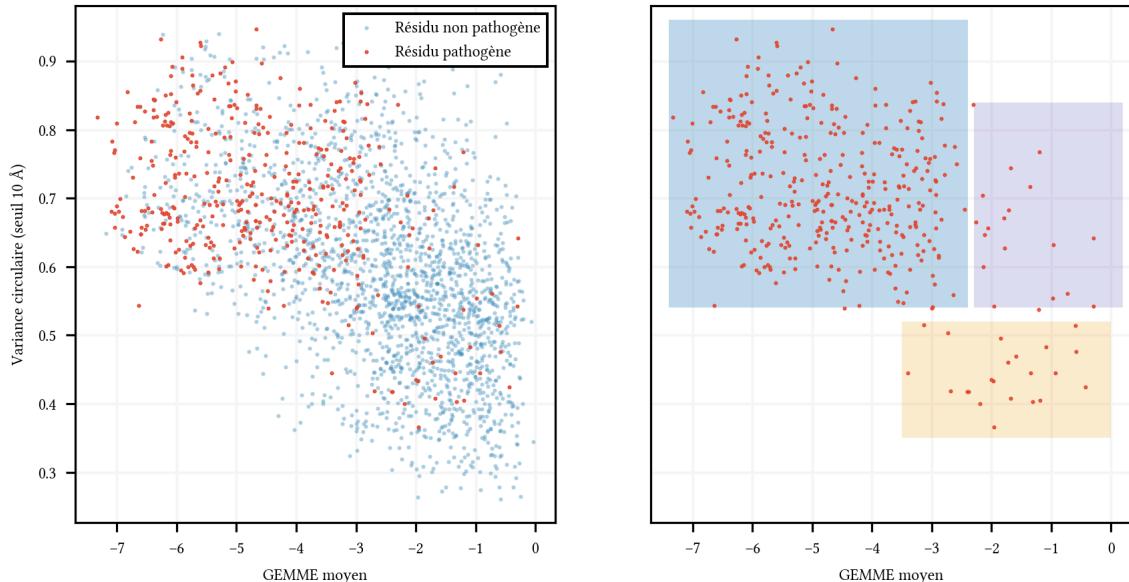


Fig. 16 Classification possible des mutations

Conclusion

Nous avons prédit la structure de la fibrilline 1 et évalué la fiabilité de cette prédiction. Sur cette base, nous avons dérivé des descripteurs structuraux pour chaque position, ainsi que des descripteurs non-structuraux.

En utilisant ces descripteurs, nous avons tenté d'effectuer une classification non supervisée des résidus de la protéine en tenant compte du contenu des positions ayant des mutations connues comme pathogènes, telles que décrites par nos collaborateurs à l'hôpital Bichat. Nous n'avons pas pu mettre en évidence de groupes bien définis de résidus à partir des descripteurs. Il est probable que différents mécanismes moléculaires soient responsables de la pathogénicité des mutations, visible notamment par l'étendue de l'exposition des résidus, la large répartition des mutations dans la protéine, ainsi que la diversité de phénotypes. La présence de mutations à des positions présentant une faible pression de sélection nous interroge aussi sur le fonctionnement de ces mutations.

Parmi les extensions possibles au projet, nous pouvons par exemple citer l'étude des épissages alternatifs ou encore l'étude des partenaires et de leur position de liaison avec la fibrilline 1.

Bibliographie

- [1] « Gene: FBN1 ». Consulté le: 21 mai 2024. [En ligne]. Disponible sur: http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000166147;r=15:48408313-48645721
- [2] « P35555 · FBN1_HUMAN ». Consulté le: 19 mai 2024. [En ligne]. Disponible sur: <https://www.uniprot.org/uniprotkb/P35555/entry>
- [3] D. Hubmacher et S. S. Apte, « Genetic and functional linkage between ADAMTS superfamily proteins and fibrillin-1: a novel mechanism influencing microfibril assembly and function », *Cellular and Molecular Life Sciences*, vol. 68, n° 19, p. 3137-3148, oct. 2011, doi: [10.1007/s00018-011-0780-9](https://doi.org/10.1007/s00018-011-0780-9).
- [4] D. Hubmacher, K. Tiedemann, et D. P. Reinhardt, « Fibrillins: From Biogenesis of Microfibrils to Signaling Functions », *Current Topics in Developmental Biology*, vol. 75. Elsevier, p. 93-123, 2006. doi: [10.1016/S0070-2153\(06\)75004-9](https://doi.org/10.1016/S0070-2153(06)75004-9).
- [5] J. Jumper *et al.*, « Highly accurate protein structure prediction with AlphaFold », *Nature*, vol. 596, n° 7873, p. 583-589, août 2021, doi: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2).
- [6] J. Abramson *et al.*, « Accurate structure prediction of biomolecular interactions with AlphaFold 3 », *Nature*, mai 2024, doi: [10.1038/s41586-024-07487-w](https://doi.org/10.1038/s41586-024-07487-w).
- [7] A. Rives *et al.*, « Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences », *Proceedings of the National Academy of Sciences*, vol. 118, n° 15, p. e2016239118, avr. 2021, doi: [10.1073/pnas.2016239118](https://doi.org/10.1073/pnas.2016239118).
- [8] B. Lee et F. Richards, « The interpretation of protein structures: Estimation of static accessibility », *Journal of Molecular Biology*, vol. 55, n° 3, p. 379–IN4, févr. 1971, doi: [10.1016/0022-2836\(71\)90324-X](https://doi.org/10.1016/0022-2836(71)90324-X).
- [9] S. Mitternacht, « FreeSASA ». [En ligne]. Disponible sur: <https://github.com/mittinatten/freesasa>
- [10] A. Shrake et J. Rupley, « Environment and exposure to solvent of protein atoms. Lysozyme and insulin », *Journal of Molecular Biology*, vol. 79, n° 2, p. 351-371, sept. 1973, doi: [10.1016/0022-2836\(73\)90011-9](https://doi.org/10.1016/0022-2836(73)90011-9).
- [11] W. Kabsch et C. Sander, « Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features », *Biopolymers*, vol. 22, n° 12, p. 2577-2637, déc. 1983, doi: [10.1002/bip.360221211](https://doi.org/10.1002/bip.360221211).
- [12] S. Minami, « PyDSSP ». Consulté le: 21 mai 2024. [En ligne]. Disponible sur: <https://github.com/ShintaroMinami/PyDSSP>
- [13] E. Laine, Y. Karami, et A. Carbone, « GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects », *Molecular Biology and Evolution*, vol. 36, n° 11, p. 2604-2619, nov. 2019, doi: [10.1093/molbev/msz179](https://doi.org/10.1093/molbev/msz179).
- [14] M. Tekpinar, L. David, T. Henry, et A. Carbone, « PRESCOTT: a population aware, epistatic and structural model accurately predicts missense effect ». Consulté le: 20 mai 2024. [En ligne]. Disponible sur: [http://medrxiv.org/lookup/doi/10.1101/2024.02.03.24302219](https://medrxiv.org/lookup/doi/10.1101/2024.02.03.24302219)

- [15] S. Chen *et al.*, « A genomic mutational constraint map using variation in 76,156 human genomes », *Nature*, vol. 625, n° 7993, p. 92-100, janv. 2024, doi: [10.1038/s41586-023-06045-0](https://doi.org/10.1038/s41586-023-06045-0).