Découverte et analyse de signatures à grande échelle dans les protéines amyloïdes.

Sandrine Pawlicki, Anne-Sophie Valin, Mathieu Giraud, Gilles Georges, Christian Delamarche, Grégory Ranchy

Découverte et analyse de signatures à grande échelle dans les protéines amyloïdes

Sandrine Pawlicki[1], Anne-Sophie Valin[2], Grégory Ranchy[2], Mathieu Giraud[2], Gilles Georges[2] et Christian Delamarche[1]

[1] Equipe Structure et Dynamique des Macromolécules, UMR CNRS 6026,
Université de Rennes I, Campus de Beaulieu, Nb 13, 35042 RENNES Cedex, France
{sandrine.pawlicki, christian.delamarche}@univ-rennes1.fr
[2] Symbiose, IRISA, INRIA, CNRS, Université de Rennes 1
Campus de Beaulieu, 35042 RENNES Cedex, France
{avalin, granchy, mgiraud, georges}@irisa.fr

Résumé:
Le terme "amyloïde" décrit des dépôts intra ou extracellulaires, principalement composés de protéines assemblées en fibres. Ces fibres amyloïdes présentent des caractéristiques particulières : structure dite "cross-beta", biréfringence verte après coloration au rouge Congo et résistance aux protéases. Pourtant, les protéines mises en évidence au sein de ces fibres appartiennent à une trentaine de familles sans ressemblance structurale ou fonctionnelle évidente, à l'exception de cette capacité à former des agrégats fibrillaires insolubles. Les fibres amyloïdes sont caractéristiques, notamment, de plusieurs pathologies neurodégénratives majeures, telle que la maladie d'Alzheimer. Toutefois, les mécanismes moléculaires conduisant à la formation des fibres amyloïdes sont encore largement inconnus. La comparaison des différentes familles de protéines amyloïdes devrait permettre de mettre en évidence des déterminants physico-chimiques impliqués dans ces mécanismes d'agrégation.
Le travail présenté dans cet article est divisé en trois points principaux :

- *La première étape a été la construction d'une base de connaissance, appelée AMYPdb, dédiée au stockage d'informations sur les familles de protéines amyloïdes et leurs signatures de séquences. AMYPdb est la première base de données consacrée à l'identification bioinformatique de signatures de séquences pouvant jouer un rôle dans l'agrégation protéique et l'assemblage en fibres.*
- *La seconde partie a été la découverte à grande échelle de signatures pour chacune des familles de protéines amyloïdes. Cela a généré 3332 motifs, qui ont ensuite été recherchés dans les 2 millions de séquences d'UniProtKB. 14 millions d'occurences de ces motifs ont ainsi été découvertes.*
- *Dans un troisième temps, nous avons analysé qualitativement chaque signature en utilisant trois critères : la sensibilité, la spécificité et la corrélation. Nous avons ainsi mis en lumière des signatures de meilleure qualité que les motifs déjà connus des protéines amyloïdes. Ces signatures ont ensuite été utilisées pour identifier de nouvelles protéines appartenant aux familles amyloïdes.*

Mots-clés: protéines amyloïdes, signature de séquence, découverte de motif, recherche de motif

1 Introduction

1.1 Les Protéines Amyloïdes et les prions

L'amyloïde est un dépôt de nature essentiellement protéique qui peut se former à l'intérieur ou à l'extérieur des cellules. Les protéines agrégées au sein de ces dépôts y sont assemblées sous forme de fibres dites amyloïdes, toujours similaires malgré la diversité des conditions d'agrégation et des

protéines qui les composent. En effet, ces protéines, aussi qualifiées d'amyloïdes, appartiennent à plus de trente familles protéiques, qui n'ont de point commun ni au plan fonctionnel ni au plan de leur structure native, et qui exercent chacune une fonction spécifique au sein des organismes qui les expriment. Dans certaines conditions (environnement cellulaire, mutations, protéolyse, ...), ces protéines montrent une capacité commune de changement conformationnel puis d'auto assemblage en agrégats fibrillaires ordonnés, l'amyloïde. Les mécanismes moléculaires de conversion des protéines normales en polymères insolubles sont encore largement inconnus [1].

Les protéines amyloïdes sont impliquées dans de nombreuses maladies qui peuvent toucher l'ensemble de l'organisme (amyloïdoses systémiques) ou bien se limiter à un organe particulier (cerveau, coeur...) [2]. Certaines protéines amyloïdes sont ainsi à l'origine de graves pathologies neurodégénératives : protéines APP et Tau impliquées dans la maladie d'Alzheimer [3]; Huntingtine liée à la maladie de Huntington [4], ... Ces pathologies ont un développement très lent et il n'existe pas, à ce jour, de traitement pour bloquer la formation des dépôts d'amyloïde. D'autres protéines amyloïdes sont à l'origine d'amyloïdoses plus bénine. Un exemple en est l'insuline, qui peut former des dépôts localisés au niveau des sites d'injection chez les diabétiques [5].

Les prions sont un cas à part parmi les protéines amyloïdes. Ce sont des particules infectieuses, de nature protéique, qui se propagent en catalysant le changement de conformation d'une forme native en une forme fibrillaire [6]. On connaît aujourd'hui une seule famille de protéines prioniques chez les mammifères : il s'agit de la protéine PrP (Prion Protein), responsable des encéphalopathies spongiformes, de l'insomnie familiale fatale, et d'autres maladies neurodégénératives. Chez les eucaryotes inférieurs (levures et champignons), on connaît plusieurs prions, mais ceux-ci ne semblent pas présenter de caractère pathogène pour l'homme.

1.2 Le Mécanisme d'Agrégation

La formation de fibres amyloïdes, comme présentée dans la figure 1, est un évènement court-circuitant le devenir normal des protéines, qu'elles soient en cours de maturation ou ayant déjà acquis leur conformation mature [7]. La déstabilisation structurale des protéines amyloïdes, et leur réorganisation dans une conformation favorisant leur agrégation, passe par l'association de petits agrégats ordonnés ou unités de nucléation. Ces éléments de nucléation possèdent la capacité de recruter de nouvelles protéines ayant déjà acquis une conformation amyloïdogénique, voire même, dans le cas des prions, de recruter des protéines normales et d'induire leur changement de conformation. La croissance des unités de nucléation donne de longs protofilaments qui s'enroulent pour former les fibres amyloïdes matures (figure 2). Des études par diffraction aux rayons X montrent qu'au sein des protofilaments les protéines sont empilées avec une disposition parallèle ou antiparallèle de feuillets betas. Ces derniers sont donc perpendiculaires à l'axe de la fibre, donnant ainsi son nom à la structure cross-beta qui caractérise toutes les fibres amyloïdes (figure 1).

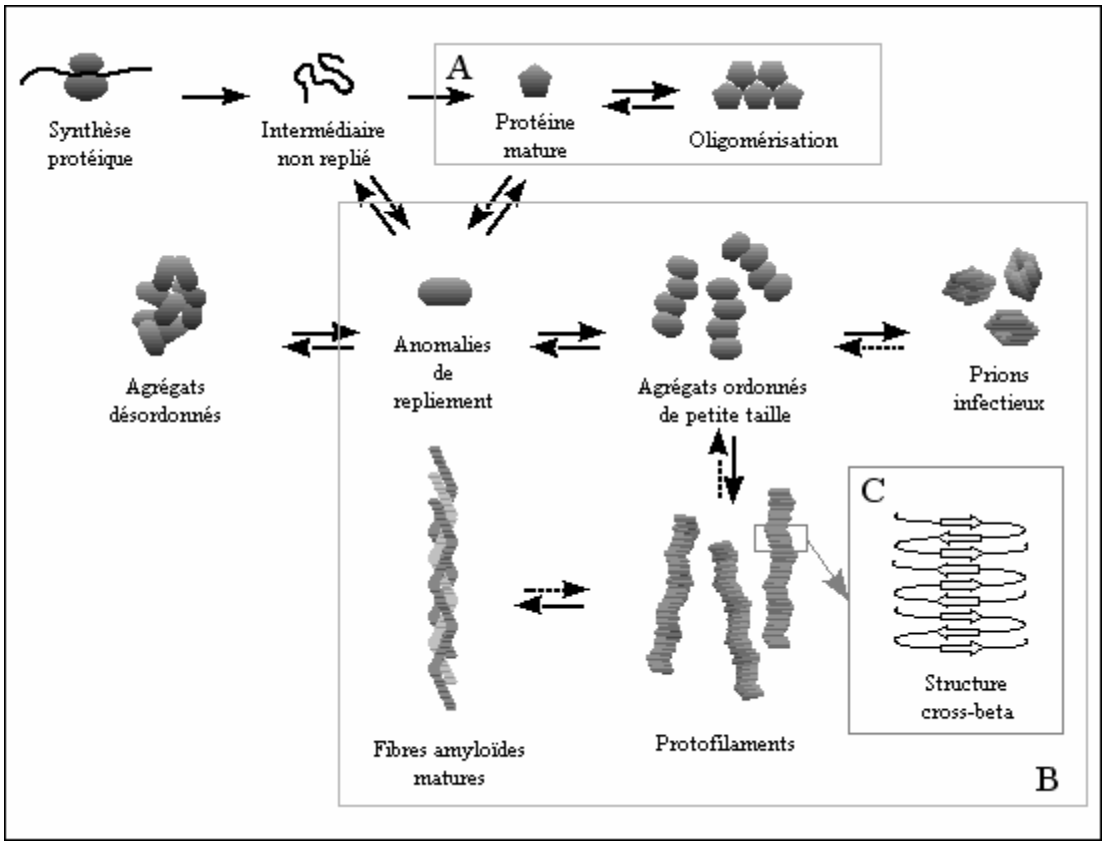

Figure 1. Mécanisme d'agrégation des protéines amyloïdes. A: Devenir normal des protéines. B: Agrégation sous forme de fibres amyloïdes suite à l'apparition d'anomalies de repliement. C: Empilement des feuillets formant la structure cross-beta.

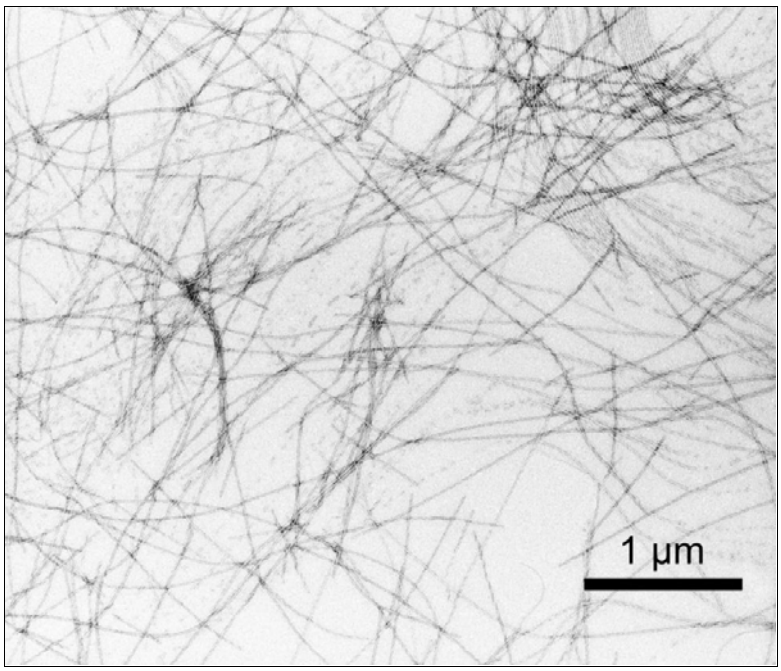

Figure 2. Assemblage en fibres amyloïdes d'un hexapeptide synthétique dérivé de la protéine Tau.

1.3 Objectifs de l'étude

Les protéines amyloïdes forment un ensemble très hétérogène, mais sont pourtant caractérisées par un même phénomène de formation de fibres insolubles dont la structure est conservée. Cette diversité structurale et fonctionnelle conduit à l'hypothèse que cette agrégation en fibres amyloïdes serait liée davantage au squelette polypeptidique, plutôt qu'à la structure ou à la fonction des protéines amyloïdes sous leur forme native [7]. Il doit donc exister des signaux dans la séquence primaire des protéines amyloïdes, qui seraient impliqués dans les mécanismes moléculaires mis en jeu lors de l'agrégation fibrillaire. Ces signaux pourraient être des biais compositionnels, des signatures, des sites de modifications post-traductionnels, des répétitions, des sites d'interactions, des sites de protéolyse, etc. Nous avons choisi de privilégier l'analyse des séquences primaires par le biais de la découverte et de l'analyse de signatures à grande échelle.

Pour cela, nous avons développé une base de connaissances dédiée aux protéines amyloïdes, appelée AMYPdb (AMYloid Protein database). Cette base de données a été conçue pour regrouper, au sein d'une même structure, l'ensemble des informations concernant les protéines amyloïdes et disponibles dans les banques de données publiques, ainsi que les résultats des découvertes et analyses de signatures conduites sur ces protéines. En effet, malgré l'importance de ces protéines dans le domaine médical, il n'existe à ce jour aucune banque de données spécialisée dédiée à cette thématique.

Nous avons donc mis au point une base de données et son interface, capable d'archiver, d'analyser et de traiter rapidement et efficacement l'ensemble des données provenant de l'étude des centaines de séquences appartenant aux différentes familles de protéines amyloïdes. Nous avons aussi travaillé avec les outils de découverte et d'analyse de motifs de la plateforme de bioinformatique OUEST-Genopole®. Nous avons ainsi adapté à notre usage des outils déjà existants, et nous en avons développé de nouveaux pour répondre à nos besoins spécifiques, notamment en terme de quantité de données à traiter.

2 Données brutes contenues dans AMYPdb

Pour réaliser la base de données sur les protéines amyloïdes, trois sources principales d'informations ont été choisies. Il s'agit de trois banques de données publiques : UniProtKB (http://www.expasy.uniprot.org/) [8], PROSITE (http://us.expasy.org/prosite/) [9] et MEDLINE (http://medline.cos.com/).

Afin de faciliter la recherche et l'intégration des données dans AMYPdb, nous avons utilisé le logiciel SRS (Sequence Retrieval System) [10], qui permet, *via* une même interface, d'interroger un grand nombre de banques de données publiques spécialisées dans le domaine des protéines, ou plus généralement de la biologie. SRS présente l'avantage de proposer la sauvegarde des résultats des requêtes sous la forme de fichiers XML (eXtended Markup Language), un format parfaitement structuré et particulièrement adapté pour l'exploitation de données textuelles, comme c'est le cas ici.

L'intégration des données contenues dans ces fichiers XML se fait en ligne, à partir de la partie administration de l'interface d'AMYPdb. Nous avons pour cela mis au point des parseurs qui traitent les fichiers XML en fonction de l'origine des données qu'ils contiennent (UniProtKB, MEDLINE, ou PROSITE).

Dans ce travail, les données de séquences et bibliographiques, provenant d'UniProtKB et de MEDLINE, ont été sélectionnées par utilisation de mots-clés décrivant les différentes familles de protéines amyloïdes. Les résultats de ces requêtes ont été triés afin de ne sélectionner que les données correspondant aux protéines d'intérêt. Ce premier ensemble d'informations a été intégré dans AMYPdb. Nous avons ainsi constitué une base de travail regroupant les données concernant plus d'un millier de séquences protéiques réparties entre 33 familles de protéines amyloïdes.

Nous avons ensuite soumis plusieurs séquences phylogénétiquement éloignées de chaque famille de protéines amyloïdes à ScanProsite (http://www.expasy.org/tools/scanprosite/) [11], un outil lié à la banque PROSITE, et qui permet entre autres de rechercher, dans une séquence protéique choisie par l'utilisateur, les motifs PROSITE qui sont présents dans cette séquence. Cela nous a permis de lister 37

identifiants de motifs PROSITE contenus dans les protéines amyloïdes. Parmi ces 37 motifs, 12 sont des motifs génériques, décrivant par exemple des sites de phosphorylation ou de glycolysation. Les 25 autres identifiants se réfèrent à des motifs caractérisant 17 des 33 familles de protéines amyloïdes. Ces 37 identifiants nous ont permis de récupérer, *via SRS*, les descriptions des motifs correspondant, au format XML, que nous avons intégré dans AMYPdb (Figure 3 A et B).

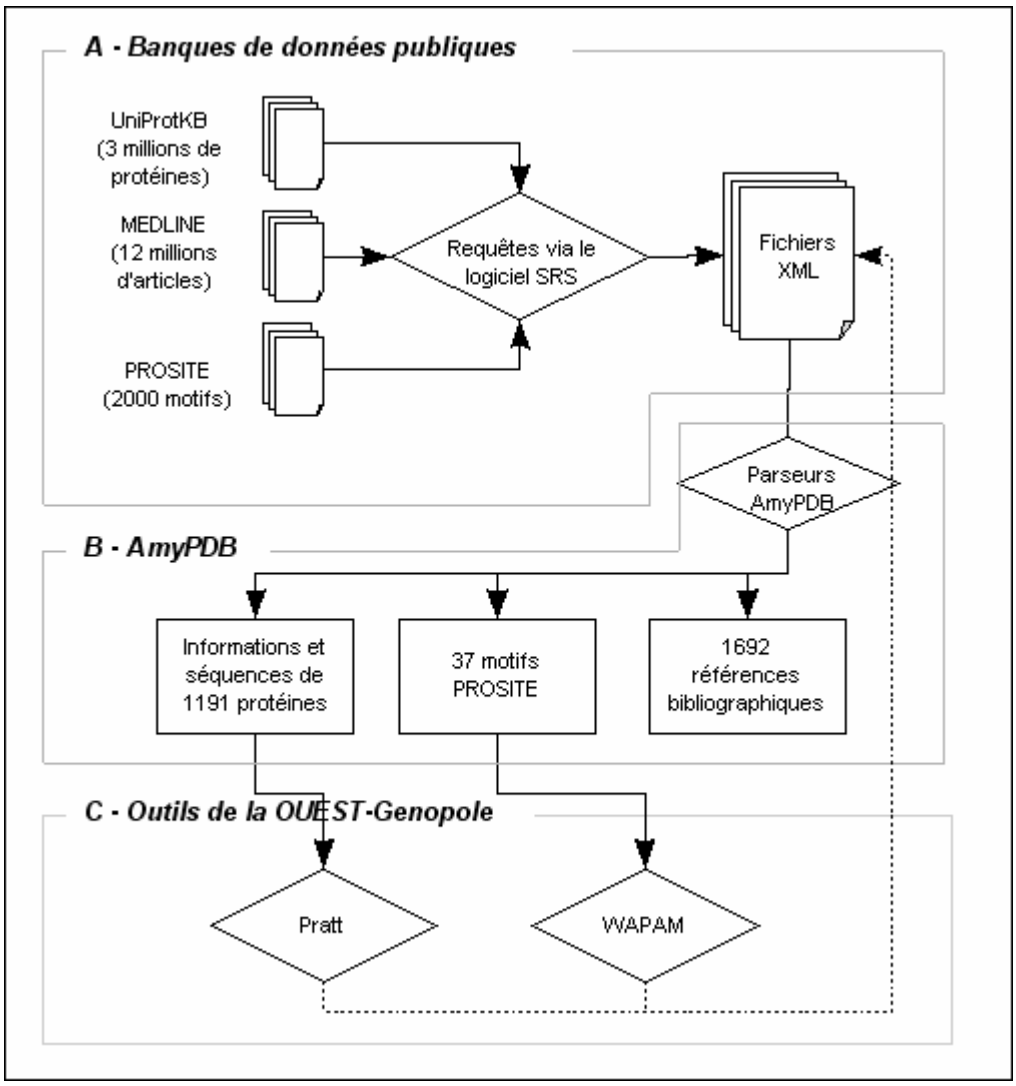

Figure 3. Stockage des informations et extraction de connaissances.

3 Extraction de connaissances

3.1 Introduction

Comme dit précédemment, nous avons fait l'hypothèse de l'existence de caractéristiques communes dans les séquences des protéines amyloïdes. Nous avons choisi d'explorer cette hypothèse par le biais d'une découverte de signatures à grande échelle. Nous avons pour cela utilisé plusieurs outils d'analyse de motifs actuellement disponibles sur la plate-forme bioinformatique de OUEST-Genopole® (http://genouest.org). Nous en avons développé et amélioré certains à cette occasion, pour répondre aux besoins spécifiques d'une telle thématique.

Les outils que nous avons utilisés sont Pratt, pour la découverte de motifs, et Wapam/Rdisk, pour

rechercher dans UniProtKB toutes les protéines contenant les motifs découverts avec Pratt. Ces deux outils proposent notamment, parmi différents formats de sortie, de récupérer les résultats sous la forme de fichiers XML. C'est ce format que nous avons exploité, en créant des parseurs spécifiques à chaque outil. Cela a beaucoup facilité la gestion de la masse des données générées, en croissance exponentielle au fur et à mesure des analyses successives (Figure 3C).

3.2 Découverte de signatures

Pour découvrir des signatures caractérisant les différentes familles de protéines amyloïdes, nous avons utilisé le logiciel Pratt [12]. Cet outil permet de découvrir automatiquement les motifs conservés dans un ensemble de séquences, de nombreux paramètres offrant la possibilité de définir la complexité et la couverture des motifs à découvrir. Un score de pertinence est donné à chacun des motifs ainsi trouvés. Une heuristique de raffinement permet alors de dégénérer les motifs en fonction des paramètres choisis par l'utilisateur. La plateforme bioinformatique de OUEST-Genopole® propose une version 2.1, modifiée afin de permettre l'utilisation de clusters d'acides aminés notamment définis par les utilisateurs.

Pratt proposait déjà l'utilisation de plusieurs clusters prédéfinis, basés sur les propriétés physico-chimiques des acides aminés : hydrophobicité, accessibilité, charge, volume, acide aminé aliphatique ou arômatique. Cette fonctionnalité a été améliorée en offrant la possibilité de définir ses propres groupes d'acides aminés. Cela permet à tout utilisateur d'orienter la découverte de motifs dans ses séquences en fonction de certaines propriétés des acides aminés importantes pour sa thématique, et non prises en compte dans les clusters par défaut de Pratt. Qu'ils soient ou non définis par l'utilisateur, l'utilisation de clusters d'acides aminés permet d'obtenir des motifs ayant un score de fitness bien plus élevés les motifs découverts sans utilisation de clusters. La figure 4 présente ainsi les scores des quinze premiers motifs découverts à partir d'un jeu de 173 séquences de la famille des protéines PrP, en utilisant (colonnes de gauche) ou non (colonnes de droite) les clusters d'acides aminés prédéfinis par Pratt.

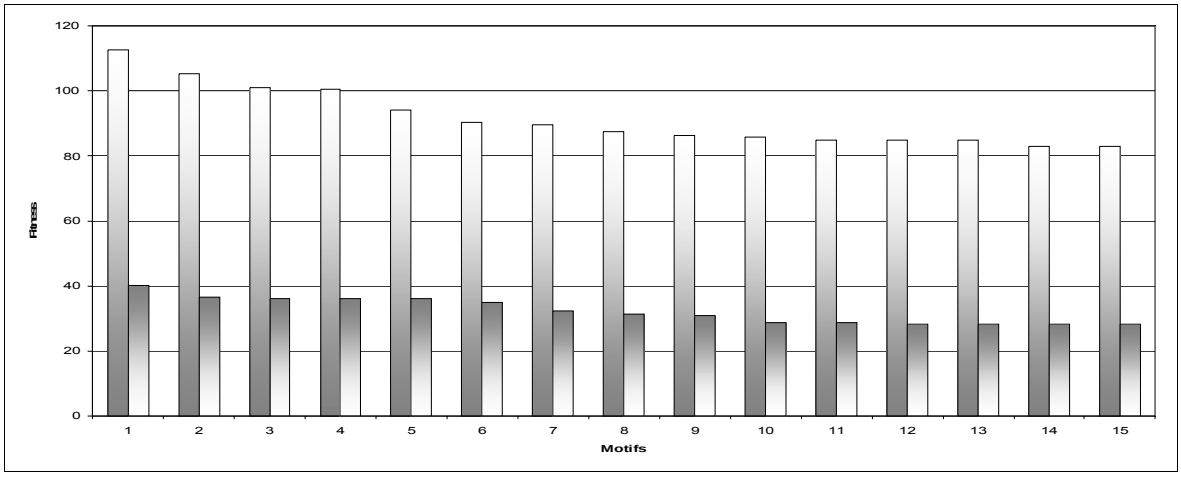

Figure 4. Qualité des 15 meilleurs motifs découverts par Pratt, avec ou sans utilisation de clusters d'acides aminés, à partir d'un groupe de séquences de la protéine prion PrP. Pour chaque numéro de motif, la colonne de gauche indique la fitness du motif découvert avec utilisation de cluster, et la colonne de droite la fitness du motif.

Le motif possédant le plus haut score a été découvert en utilisant les clusters :

G-x(0,1)-A-A-A-x(0,1)-G-A-[IV]-[GV]-x(2)-[ILM]-[GLS]-G-[INY]-[ALMV]-[LMV]-G-[RS]-x(3)-[GHQR]-[FMP]-x-[ILMNVY]-x-[FLM]-[DEG]-x-[DEPR]-x-[EY]-[DNRSY]-x-[WY]-[WY]-x-[EQ]-[MN]-[MPQS]-x-R-[VY]-[PY]-[ENRS]-[PQR]-[MV]-[MY]

Ce motif est beaucoup plus riche que le motif découvert sans prendre en compte les clusters d'acides

aminés, présenté ci-dessous :

K-x(2,3)-K-x(3,4)-G-A-A-A-x(0,1)-G-A-x(6)-G-x(3)-G

Pour chaque famille de protéines amyloïdes disposant de suffisamment de séquences, nous avons défini un ou plusieurs groupes à soumettre à Pratt. Pour cela, nous avons sélectionné les séquences les plus divergentes possibles, de façon à avoir au minimum une dizaine de séquences par groupe. Nous avons alors traité ces groupes de séquences avec Pratt, avec plusieurs séries de paramètres pour chaque groupe, en recherchant des motifs spécifiques à 100% des séquences de chaque groupe.

Nous avons ainsi cherché à découvrir des motifs d'une longueur de 10 à 20 acides aminés, en ne sélectionnant que les 20 meilleurs motifs dans chaque cas. Cette taille de motifs devrait permettre d'obtenir des signatures ni trop lâches ni trop stringentes. L'utilisation des clusters d'acides aminés permettant d'obtenir les meilleurs motifs, une première série de motifs a été découverte en utilisant les clusters prédéfinis et présélectionnés par défaut. Une seconde série de motifs a été générée en utilisants des clusters définis, après analyse bibliographique, en fonction de critères liés à la formation d'amyloïde. L'un de ces groupes de clusters comprend ainsi trois ensembles d'acides aminés selon leur propension à former des brins beta (CIFTWYV), des hélices alpha (AREQLK), ou des régions autres (NDGHMPS) [13]. Un autre groupe contient deux ensembles d'acides aminés, selon leur propension à se retrouver (CHILMFWYV) ou non (ARNDEQGKPST) au sein d'interfaces entre protéines [14,15]. En plus de l'utilisation de clusters, nous avons aussi fait varier le paramètre RG (on ou off), qui permet d'augmenter la liste des symboles ambigüs, et donc découvrir des motifs plus relâchés.

L'utilisation de Pratt avec l'ensemble de ces jeux de séquences et de paramètres nous a permis de générer près de 3300 motifs décrivant les différentes familles de protéines amyloïdes.

3.3 Recherche de motifs

L'ensemble des motifs contenus dans notre base (motifs déjà connus ou inférés via les méthodes expliquées dans les parties précédentes) a été recherché dans UniprotKB. Pour ces recherches, nous avons utilisé Wapam (Weighted Automaton Pattern Matching, recherche de motifs par automates pondérés), un outil développé dans l'équipe Symbiose et accessible sur la plate-forme bioinformatique de OUEST-Genopole®.

L'interface Web de Wapam donne aux utilisateurs la possibilité de lancer leur requête sur les serveurs de la plate-forme (SunFire 6800), mais aussi sur l'accélérateur Rdisk, une architecture spécialisée constituée de cartes avec des processeurs reconfigurables FPGA qui accélèrent les recherches [16]. Le logiciel Wapam transforme d'abord les motifs sous forme d'automates pondérés. Ces automates sont créés directement à partir de motifs PROSITE (figure 5) et peuvent être améliorés manuellement. Les poids sont comptabilisés le long des transitions reconnaissant les motifs, et le poids final se compare à un seuil : il est ainsi possible de fixer une limite quelconque aux erreurs de substitution.

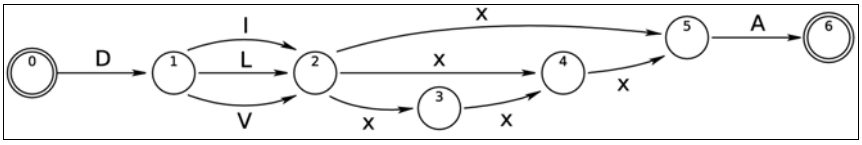

Figure 5 Automate pondéré représentant le motif D-[ILV]-x(1,3)-A. L'automate est pondéré, c'est-à-dire qu'un poids est affecté à chaque transition : pour compter les erreurs de substitution, chaque transition a un poids de 0 lorsqu'elle est réussie et de -1 sinon. Le motif est reconnu lorsque l'état final (6) est actif avec un score supérieur au seuil d'erreur fixé. Il est aussi possible de calculer des scores plus complexes.

Dans la version logicielle, tous les états de l'automate sont émulés séquentiellement. Dans la version matérielle (Rdisk, figure 6), les automates sont directement "câblés" sur les processeurs reconfigurables, ce qui permet une évaluation simultanée des états. Ce cablage utilise autant d'éléments matériels que de transitions et d'états dans l'automate [17]. Les processeurs utilisés ont une surface pouvant câbler des automates ayant jusqu'à une centaine de transitions. Enfin, 30 cartes se

partagent le "scan" de la banque UniprotKB (1/30ème de la banque par carte). L'ensemble du prototype Rdisk a été conçu pour filtrer rapidement les bases de données, les disques durs étant directement reliés aux processeurs FPGA:

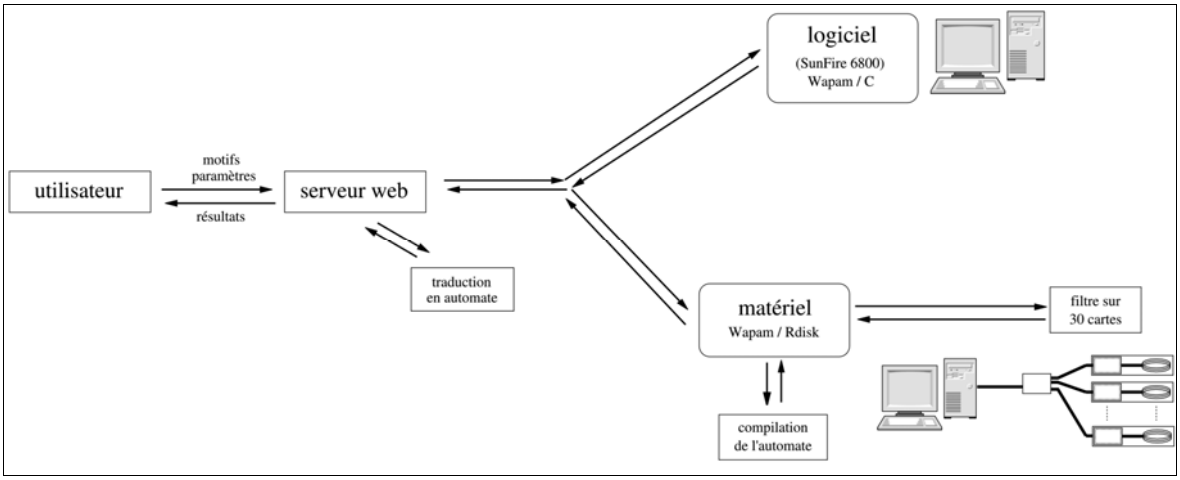

Figure 6. Utilisation de Wapam sur la plateforme de OUEST-Genopole®.

Le tableau 1 compare les temps d'exécution des recherches de motifs selon l'implémentation utilisée. Pour ne pas surcharger les serveurs, la recherche peut être arrêtée dès qu'il y a plus de 2000 résultats (auto-stop). L'accélération apportée par Rdisk est encore plus importante à partir du deuxième lancement, lorsque les motifs ont déjà été compilés, car Wapam / Rdisk se souvient des automates pondérés compilés précédemment. La modification du seuil d'erreur ne demande pas une nouvelle compilation.

	Wapam logiciel	Wapam + auto-stop	Wapam / Rdisk	Wapam / Rdisk + précompilation
un motif	2605 s	2003 s	72 s	23 s
tous les motifs	100 jours*	77 jours*	< 3 jours	< 1 jour

Tableau 1. Comparaison des temps de recherche de motif entre l'implémentation logicielle de Wapam et l'accélération matérielle Wapam / Rdisk.(moyenne sur 50 motifs pris aléatoirement parmi les 3331) *estimation

Afin de pouvoir évaluer la qualité des motifs précédemment découverts et d'identifier de nouvelles séquences appartenant aux différentes familles de protéines amyloïdes, nous avons utilisé Wapam pour rechercher toutes les protéines référencées dans UniProtKB et contenant ces motifs. Nous avons ainsi récupéré plus de 268 000 séquences, un nombre élevé dû à la présence parmi nos 3300 motifs de motifs très peu stringents, et donc contenus dans un grand nombre de protéines. La limite de 5000 séquences, imposée par Wapam, a toutefois permis de ne pas récupérer l'ensemble des protéines contenant ces motifs peu stringents, et qui présentent peu d'intérêt pour notre étude.

4 Analyse des motifs dans AMYPdb

4.1 Synthèse des données sur les motifs

Les approches successives de découverte, comparaison et recherche de motifs, puis l'intégration des résultats obtenus dans AMYPdb ont généré près de 400 000 enregistrements, inexploitables en l'état. Nous avons donc procédé à plusieurs séries de calculs analytiques afin de trier, corréler, synthétiser et rendre accessible cet ensemble de données. L'évolution de la masse de données générée par les

différents outils et analysée dans AMYPdb est décrite dans le tableau 2.

Etape	Jeu de données à analyser	Résultats des analyses
Création d'AMYPdb	Références bibliographiques	33 familles, 1191 séquences, 37 motifs
Découverte de motifs (Pratt)	46 jeux de séquences, 6 jeux de paramètres	3295 motifs
Recherche de motifs (WAPAM)	3332 motifs, 2 034 540 séquences contenues dans UniProtKB	268 663 séquences
Analyse des motifs	3332 motifs, 268 663 séquences	14 430 964 occurences, 13 902 descriptions motif/famille
Données totales d'AMYPdb	14 839 239 enregistrements et 1,3GB de données	

Tableau 2. Evolution de la quantité de données stockées dans AMYPdb en fonction des analyses successives.

Nous avons commencé par répertorier l'ensemble des occurences, et des positions de ces occurences, de tous les motifs dans les proteines amyloïdes et non amyloïdes. Nous avons pour cela utilisé les fonctions liées aux expressions régulières proposées par PHP et MySQL. Le script que nous avons ainsi créé a permis d'identifier, en une semaine de calculs, près de 14,5 millions d'occurences des 3300 motifs issus de Pratt et PROSITE dans les 250 000 séquences stockées dans AMYPdb, atteignant une taille de stockage de plus d'1 GB. Parmi ces 14,5 millions d'occurences, plus de 450 000 concernent des protéines amyloïdes.

Afin de pouvoir manipuler, au niveau des familles protéiques, l'information liée à toutes ces occurences dans les séquences, nous avons calculé, pour chaque couple motif/famille de protéines amyloïdes, les données statistiques permettant d'évaluer la qualité du motif pour cette famille protéique. Nous avons privilégié l'utilisation de trois critères couramment utilisés pour décrire la qualité d'un motif : la sensibilité, la sélectivité, et le facteur de corrélation [18]. Ces trois critères reposent sur la définition stricte de quatre ensembles de données, les vrais et faux positifs, et les vrais et faux négatifs. Pour notre thématique des protéines amyloïdes, les données composant ces quatre ensembles sont décrites dans le tableau 3. Concernant les protéines n'appartenant pas à la famille, le nombre total de ces protéines correspond au nombre de séquences d'UniProtKB lors de la recherche avec WAPAM. Cette recherche a été effectuée sur les versions 48.4 de SwissProt et 30.4 de TrEMBL, qui répertoriaient 2 034 540 protéines.

	Protéines de la famille	Protéines n'appartenant pas à la famille
Contient le motif	Vrais Positifs	Faux Positifs
Ne contient pas le motif	Faux Négatifs	Vrais Négatifs

Tableau 3. Définition des quatre ensembles de séquences nécessaires au calcul de la sensibilité, de la sélectivité et du facteur de corrélation de chaque motif.

Pour chaque couple motif/famille, nous avons donc calculé le nombre de vrais et faux positifs et de vrais et faux négatifs, avant de calculer la valeur des trois critères choisis.

Le premier de ces critères est la sensibilité d'un motif, c'est-à-dire sa capacité à récupérer toutes les protéines de la famille qu'il décrit. Cette sensibilité correspond à la proportion des séquences de la famille contenant le motif d'intérêt. Sa valeur est calculée à partir de la formule suivante :

$$Sensibilité = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}}$$

Le second critière que nous avons choisi est la sélectivité, qui est la capacité d'un motif à ne récupérer que les protéines de la famille considérée. Il s'agit en pratique de la proportion de protéines n'appartenant pas à la famille et ne contenant pas le motif. La formule permettant de calculer ce paramètre est la suivante :

$$Sélectivité = \frac{\text{Vrais Négatifs}}{\text{Vrais Négatifs} + \text{Faux Positifs}}$$

Afin de pouvoir trier encore plus facilement nos motifs en fonction de leur qualité, nous avons utilisé un autre critère, le facteur de corrélation, calculé selon la formule présentée ci-dessous:

$$Corrélation = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP + FP)(FP + VN)(VN + FN)(FN + VP)}}$$

Ce paramètre, en plus de prendre en compte la sensibilité et la sélectivité d'un motif, va aussi varier en fonction du nombre de protéines composant la famille considérée. Ainsi, un motif ayant une sensibilité et une sélectivité élevées pourra avoir un faible coefficient de corrélation si très peu de séquences composent cette famille.

L'utilisation de ces trois paramètres a permis de synthétiser, pour chaque famille, l'information contenue dans les 14,5 millions d'occurrences de nos motifs dans les protéines amyloïdes et de la réduire à moins de 14 000 enregistrements décrivant chaque couple motif/famille.

4.2 Nouveaux motifs décrivant les familles de protéines amyloïdes et enrichissement d'AMYPdb

Toutes les informations stockées dans la base de données sont accessible (identifiant : user et mot de passe : test_user) en ligne (http://AMYPdb.univ-rennes1.fr/) par l'intermédiaire d'une interface réalisée en PHP (PHP: Hypertext Preprocessor). Cette interface permet d'accéder à toutes les données concernant chacune des familles de protéines amyloïdes, que ce soit les descriptions de ces familles, les listes des séquences, ou celles des motifs. Ces derniers sont pré-triés en fonction notamment des facteurs de corrélation, ce qui permet par exemple d'accéder très facilement aux motifs décrivant le mieux telle ou telle famille de protéines amyloïdes, ou commun à plusieurs familles.

Une première exploitation de ces données concernant les motifs a permis de lister, pour chaque famille de protéines amyloïdes, le motif caractérisant le mieux cette famille, et de comparer la qualité de ce motif à celle du meilleur motif PROSITE déjà connu pour cette famille, lorsqu'il en existait au moins un. Trois résultats de cette étude sont présentés dans le tableau 4.

Famille	Origine	Motif		
		Corrélation	Sensibilité	Sélectivité
PrP	Pratt	0,822851	0,983516	0,99996
		A-x(0,1)-A-x(0,1)-G-x(0,1)-A-[AIV]-[AGV]-[GKY]-x-[AILMV]-x-[DGR]-x(2)-[LMR]-[GPS]-[HRS]		
	PROSITE	0,816184	0,945055	0,999965
		E-x-[ED]-x-K-[LIVM](2)-x-[KR]-[LIVM](2)-x-[QE]-M-C-x(2)-Q-Y		
Tau	Pratt	0,797723	1	0,999998
		G-S-x(0,1)-D-N-[IMV]-[KNRT]-H-x-P-G-G-G-[EKNS]-[KV]-[KQ]-I-x-[DHTY]		
	PROSITE	0,356749	1	0,999976
		G-S-x(2)-N-x(2)-H-x-[PA]-[AG]-G(2)		
Prolactine	Pratt	0,899732	1	0,999992
		R-D-S-x-K-[IV]-[DK]-[NST]-[FY]-L		
	PROSITE	0,382435	0,926471	0,999835
		C-x-[STN]-x(2)-[LIVMFYS]-x-[LIVMSTA]-P-x(5)-[TALIV]-x(7)-[LIVMFY]-x(6) -[LIVMFY]-x(2)-[STACV]-W]		

Tableau 4. Comparaison des motifs PROSITE et Pratt pour trois familles de protéines amyloïdes.

Nous avons ensuite utilisé ces motifs représentatifs des différentes familles pour rechercher, parmi les résultats renvoyés par WAPAM, les protéines amyloïdes qui n'étaient pas encore répertoriées dans AMYPdb. Nous avons ainsi ajouté à la plupart des familles de protéines amyloïdes de nouvelles séquences, qui correspondaient soit à des protéines amyloïdes déjà annotées mais rajoutées dans UniprotKB après la création d'AmyPB, soit à des protéines non annotées et possédant les motifs caractéristiques d'une des familles de protéines amyloïdes. Cette opération a donc permis de mettre à jour notre catalogue de séquences.

A la suite de cette mise à jour, nous avons recalculé les paramètres de sensibilité, sélectivité et corrélation pour chaque couple motif/famille de protéines amyloïdes. En effet, les effectifs correspondant aux vrais et faux positifs et vrais et faux négatifs ont été modifiés par l'ajout des nouvelles séquences, et nous avons donc réévalué les valeurs des trois critères de qualité caractérisant nos motifs. L'ajout des nouvelles séquences ayant augmenté le nombre de vrais positifs et diminué le nombre de faux positifs, les nouvelles valeurs des facteurs de corrélation ainsi obtenues sont légèrement supérieures à celles calculées précédemment (Tableau 5).

Famille	Après création d'AMYPdb				Après mise à jour d'AMYPdb			
	Nb. prot.	C.	Sen.	Sel.	Nb. prot.	C.	Sen.	Sel.
PrP	272	0,822851	0,983516	0,99996	352	0,989563	0,9875	0,999999
Tau	32	0,797723	1	0,999998	36	0,942809	1	1
Prolactine	92	0,899732	1	0,999992	112	0,994302	1	1

Tableau 5. Evolution du nombre de protéines de quelques familles de protéines amyloïdes, ainsi que des facteur de corrélation (C), sensibilité (Sen) et sélectivité (Sel) du meilleur motif caractérisant chacune des familles.

Remerciements

Cette étude a bénéficié du support financié de la Région Bretagne (allocation de thèse de Sandrine Pawlicki). Les auteurs remercient le personnel de la plate forme bioinformatique de la OUEST-Genopole® pour leur intérêt pour ce projet et pour la mise à disposition des machines.

Réferences

[1] C. Lee et M.-H. Yu, Protein folding and diseases. J Biochem Mol Biol., 38(3):275-280, 2005.
[2] D.J. Selkoe, Protein folding in fatal ways. Nature, 426(6968):900-904, 2003.
[3] M. Morishima-Kawashima et Y. Ihara, Alzheimer's Disease : β-amyloid protein and Tau. J Neurosci Res., 70:392-401, 2002.
[4] C. Landles et G.P. Bates, Huntingtin and the molecular pathogenesis of Huntington's disease. Fourth in molecular medicine review series. EMBO Rep., 5(10):958-963, 2004.
[5] F.E. Dische, C. Wernstedt, G.T. Westermark, P. Westermark, M.B. Pepys, J.A. Rennie, S.G. Gilbey et P.J. Watkins, Insulin as an amyloid-fibril protein at sites of repeated insulin injections in a diabetic patient. Diabetologia, 31(3):158-161, 1998.
[6] S.B. Prusiner, Prions. Proc Natl Acad Sci USA., 95(23):13363-13383, 1998.
[7] C.M. Dobson, Protein folding and misfolding. Nature, 18;426(6968):884-890, 2003.
[8] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi et L.S. Yeh, The Universal Protein Resource (UniProt). Nucleic Acids Res., 33: D154-159, 2005.
[9] N. Hulo, C.J. Sigrist, V. Le Saux, P.S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, et A. Bairoch. Recent improvements to the PROSITE database. Nucl Acids Res., 32:D134–D137, 2004.
[10] E.M. Zdobnov, R. Lopez, R. Apweiler et T. Etzold, The EBI SRS server – recent developments. Bioinformatics, 18:368-373, 2002.
[11] A. Gattiker, E. Gasteiger et A. Bairoch, ScanProsite: a reference implementation of a PROSITE scanning tool. Applied Bioinformatics, 1:107-108, 2002.
[12] I. Jonassen, J. Collins et D. Higgins, Finding flexible patterns in unaligned protein sequences. Protein science, 4(8):1587-1595, 1995.
[13] Y. Kallberg, M. Gustafsson, B. Persson, J. Thyberg et J. Johansson, Prediction of amyloid fibril-forming proteins. J Biol Chem., 276(16):12945-12950, 2001.
[14] P. Chakrabarti et J. Janin, Dissecting protein-protein recognition sites. Proteins, 47:334-343, 2002.
[15] R.P. Bahadur, P. Chakrabarti, F. Rodier et J. Janin, Dissecting subunit interfaces in homodimeric proteins. Proteins, 53:708-719, 2003
[16] S. Guyetant, M. Giraud, L. L'Hours, S. Derrien, S. Rubini, D. Lavenier et F. Raimbault, Cluster of re-configurable nodes for scanning large genomic banks, Parallel Computing, 31(1), pp. 73-96, 2005.
[17] M. Giraud, S. Guyetan et, D. Lavenier, Encodage linéaire d'automates pondérés: Filtrage de motifs génomiques et application sur l'architecture prototype Rdisk, TSI, 24/6, pp. 703-724, 2005.
[18] A. Brazma, I. Jonassen, I. Eidhammer et D. Gilbert, Approaches to the automatic discovery of patterns in biosequences, Reports in Informatics, 1995.