N'DEBI Mélissa - LAMOTHE Gilles - DUBRUNFAUT Lucie
Master 2 Biologie Informatique Bioinformatique

Conception d'un système d'information dédié à l'analyse de structure secondaire de type polyproline II dans les protéines

I. Introduction

L'étude des protéines passe par la caractérisation de leur séquence mais aussi par l'identification du repliement de celles-ci qui leur donne souvent des propriétés fonctionnelles spécifiques. La conformation tridimensionnelle des protéines est caractérisée par le repliement de motifs structurés et répétés que l'on appelle des structures secondaires. Certaines structures secondaires telles que les hélices alpha et les brins bêta sont fréquentes et très bien décrites. En revanche, d'autres structures secondaires sont beaucoup moins évidentes et leur découverte est beaucoup plus récente. C'est notamment le cas des polyprolines II (PPII). Lors de leur découverte, il a été observé que certaines séquences contenant des prolines successives se replient d'une manière particulières - avec des angles phi et psi particuliers - afin de donner une structure secondaire que l'on a nommé PPII. Il s'est ensuite avéré que les PPII sont en fait très souvent constituées d'acides aminés autres que la proline ; les PPII sont donc plus fréquentes que l'on ne le pensait initialement et on s'intéresse désormais à mieux identifier la présence de ces structures secondaires.

Il existe plusieurs méthodes de prédiction de structures secondaires, mais leurs prédictions peuvent différer et c'est souvent le cas pour l'assignation des PPII. Il est donc intéressant de comparer l'assignation de structures secondaires - notamment des PPII - provenant de différentes méthodes afin d'obtenir une assignation consensus. L'objectif de ce projet est de créer une base de données contenant des structures cristallographiques à haute résolution et sans redondance de séquence, ainsi que leurs assignations de structures secondaires obtenues à partir de plusieurs méthodes de prédiction (ex. PROSS, DSSP). Puis, à partir d'une interface web, l'utilisateur pourra solliciter cette base de données et visualiser des alignements de structures secondaires prédites par différentes méthodes afin d'en déduire des assignations consensus et de notamment mieux se rendre compte de la présence des PPII dans les protéines.

II. Matériel et Méthodes

A. Production des données

Afin de produire les données nécessaires à la création de la base de données, nous avons d'abord exporté - à partir de la Protein DataBank - des structures cristallographiques de protéines ayant une forte résolution (< 1.6 A) et sans redondance (ID de séquence < 20%). Pour la sélection de ces structures, nous avons utilisé l'outil PICSES développé par le laboratoire du Pr. Roland Dunbrack (ttp://dunbrack.fccc.edu/PISCES.php) ; 2979 structures ont été sélectionnées et extraites. Nous avons ensuite prédit les structures secondaires de chacune de ces structures protéiques en utilisant deux méthodes de prédiction: DSSP et PROSS. Afin d'automatiser le téléchargement et la prédiction DSSP/PROSS d'un grand nombre de structures PDB, des scripts Bash et Python ont été créés. À partir des fichiers PDB et des fichiers de sortie de DSSP et PROSS, nous avons extrait les données pertinentes à notre base de données à l'aide de scripts Python que nous avons développés. Ces informations incluent : l'identifiant PDB, la chaîne de la structure PDB, l'entête du fichier PBD, le titre du fichier PDB, la numérotation de la séquence protéique, la séquence en acides aminés, la taille de la séquence, la prédiction de structures secondaires (de DSSP et PROS), les angles Phi/Psi et la résolution cristallographique.

B. Création et peuplement de la base de données

Une fois les données générées, nous avons conçu la base de données. Le schéma relationnel est présenté dans la *figure 1*. Les différentes informations concernant les protéines, les fichiers PDB et les assignations structurales ont été intégrées à la base de données.

Nous avons décidé d'implémenter cette base de donnée en SQLite grâce au package sqlite3 disponible dans la conception de base de Python. Le script "dbCreation.py" présent dans le dossier "source" permet de créer la base de données. Elle sera générée dans un fichier nommé "database.db" à la racine du projet.

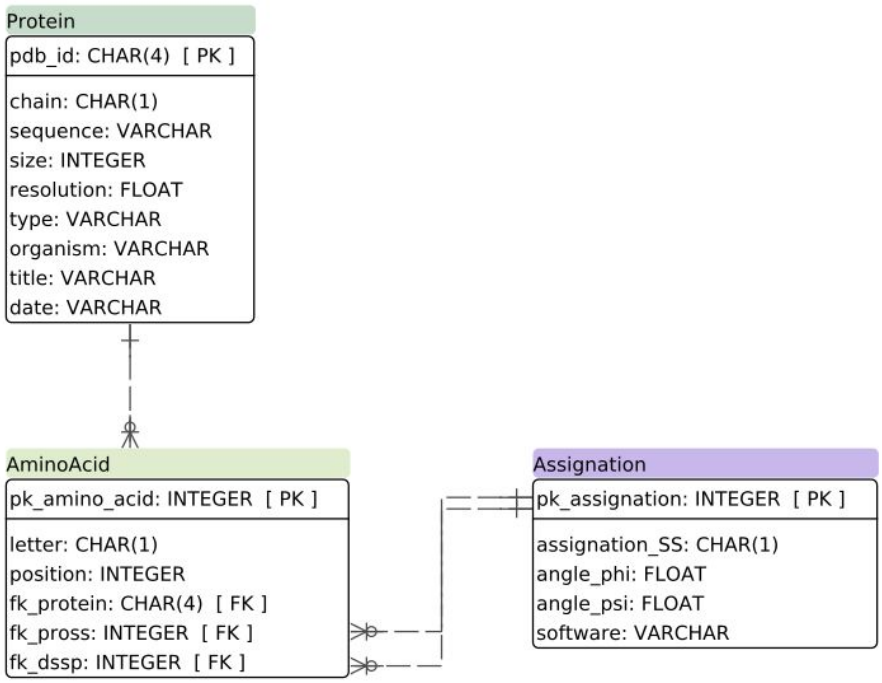

Figure 1 : Schéma relationnel de la base de données

Nous avons ensuite peuplé cette base de données. Pour cela, nous avons utilisé les fichiers générés comme décrit dans la partie précédente ainsi qu'un script python pour les parcourir (présent dans le dossier "source" et nommé "parse_assignement.py"). Les noms scientifiques des organismes d'origines des protéines ainsi que le titre des PDBs et leur date de création ont été rajoutées ultérieurement, grâce au script "parse_header.py".

C. Conception de l'interface web

Nous avons utilisé le module Flask de python qui est un framework web permettant la mise en place de l'application web de notre projet. Pour l'interaction entre le serveur web et la base de données, nous avons utilisé le module sqlite3 de python. Les pages web de l'application sont au format html et utilisent le format CSS pour la mise en page, ainsi que le Javascript pour l'interaction des pages et pour finir les templates ont été générés par Jinja2. L'arborescence des fichiers de l'application doit être respectée pour que l'application web soit utilisable (plus d'information dans le fichier README).

III. Résultats

Nous avons généré les données de 2979 structures cristallographiques de protéines, et nous avons insérée les premières 400 structures afin de pouvoir démontrer le fonctionnement de notre base de données et son interface web. Sur la page "Search" du site web, vous pouvez effectuer des requêtes de structures par code PDB (Search by PDB ID), par mots clefs (Search by Keywords) ou par autres critères (Display PDB in Database) tels que la résolution maximum/minimum et la taille de séquence maximum/minimum. La recherche par code PDB et par mot clés autorise plusieurs mots qui doivent être séparés par des espaces ou des sauts de ligne pour être reconnus. La recherche par mot clé peut se faire selon deux méthodes, "OR" qui permet de rechercher pour plusieurs mots clés, les titre de fichier PDB qui contiennent l'un ou l'autre des mots indiqués. La méthode "AND" en revanche permet la recherche de plusieurs mots clés qui doivent tous être présents dans la le titre du fichier PDB pour que celui-ci soit sélectionné en résultat de recherche.
Une fois votre requête effectuée, vous serez redirigé vers la page de résultats avec une liste de structures PDB correspondant à vos critères de recherche. Dans la *figure 3* vous pouvez voir un exemple de page de résultats. Cette page donne aussi un petit descriptif de chaque structure: PDB ID, chaîne, titre, longueur de séquence, résolution cristallographique, nombre et pourcentage de PPI selon la méthode DSSP, ainsi que le nombre et pourcentage de PPI selon la méthode PROSS.

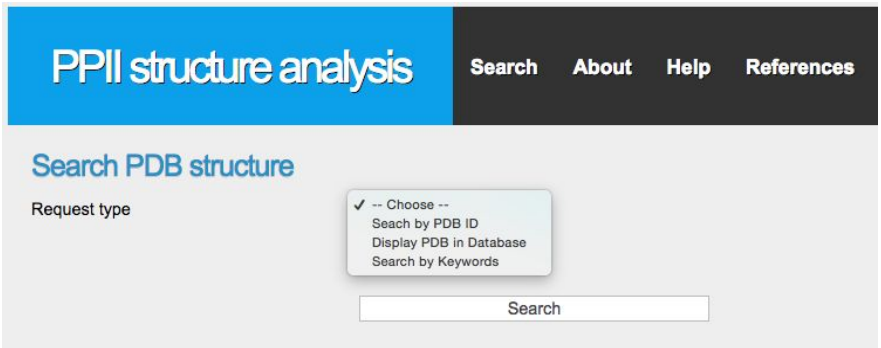

Figure 2 : Capture de la page recherche où trois différents modes de requête peuvent être sélectionnés.

Figure 3 : Capture de la page de résultat de recherche

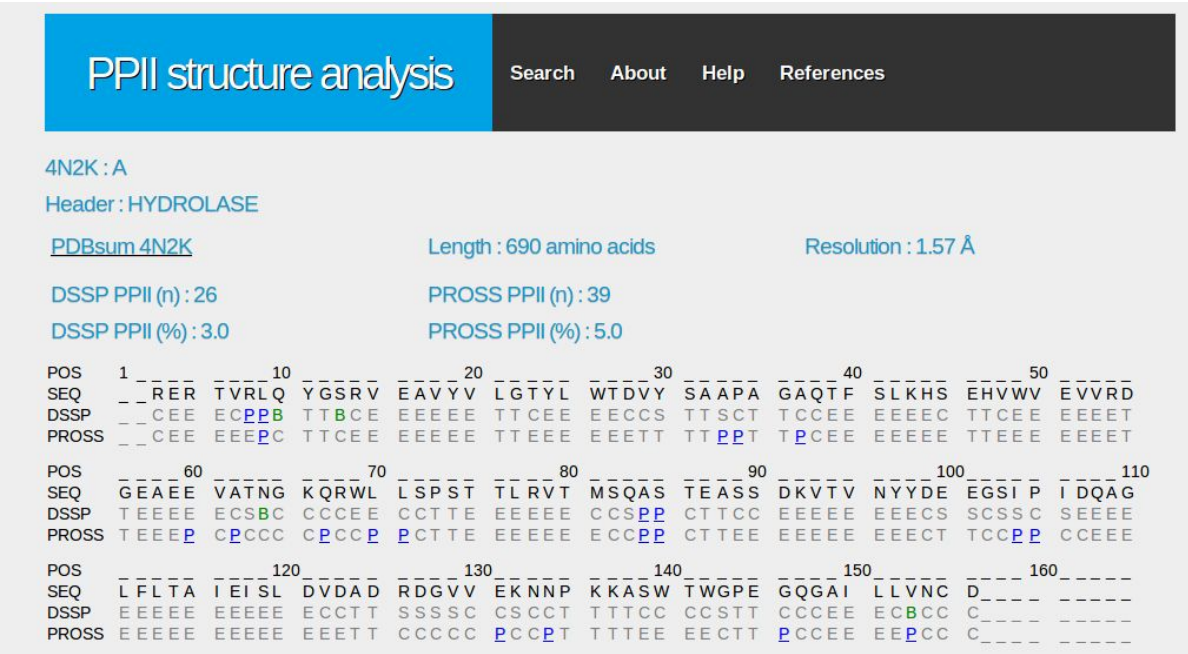

Figure 4 : Capture de la page de profil de protéine contenant un alignement des deux méthodes d'assignation de structures secondaires.

Pour plus de détails sur une protéine, une page descriptive est disponible en cliquant sur le code PDB de la protéine. La *figure 4* présente une page de profil de protéine, avec notamment un alignement entre la séquence de la protéine et les assignations de structure secondaire réalisés par les méthodes PROSS et DSSP. Un lien vers la page PDBsum du PDB est également disponible sur cette page. Il faut noter que pour certaines positions de la protéine, aucune structure n'est prédite, dans ce cas l'alignement affiche un "_" .

IV. Conclusion et Discussion

Ce projet comprend une base de données regroupant des informations sur une protéine et sa structure, mais aussi des assignations de structures secondaires. Des requêtes sur cette base de données sont possible, à la fois par un numéro PDB, par mots clés ou encore avec une sélection de paramètres (taille, résolution …). Une page d'aide est mise à disposition pour les recherches. De plus, quelques statistiques sur la base de données ont été réalisé. Les références qui nous ont permis de faire ce projet sont indiqué dans la page référence.

Plusieurs extensions sont possibles. Il aurait été intéressant de rajouter la génération des cartes de Ramachandran afin de visualiser les angles Phi et Psi associés aux prédictions de structures secondaires. D'autres améliorations visuelles peuvent êtres implémentées, par exemple l'utilisation de tableaux dynamiques pour l'affichage des séquences et des alignements, ou l'affichage d'une image de la structure protéique, voire une interaction avec la structure via une application web de visualisation moléculaire tel que JSmol. De plus certaines améliorations de recherche auraient été possibles comme la suggestion de mots clés de recherche.