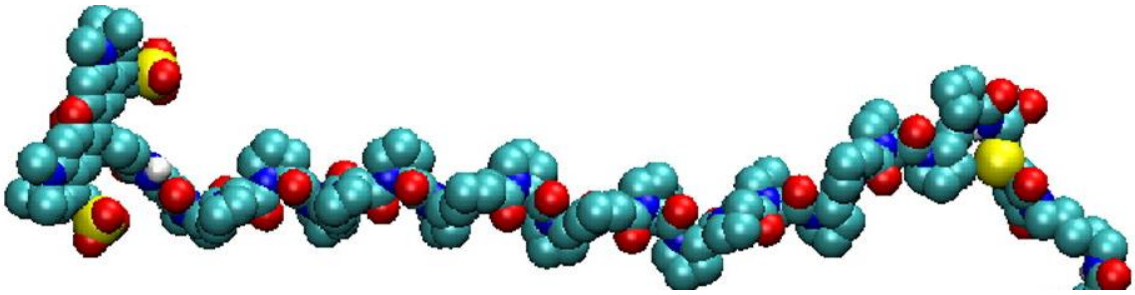
Projet Base de données : Conception d'un système d'information dédié à l'analyse de structure secondaire de type polyproline II dans les protéines

Master 2 Biologie-Informatique – Janvier 2017

Romain Coppée – Costas Bouyioukos

(projet basé sur les travaux de Alexandre G. de Brevern & Jean-Christophe Gelly)



Introduction :

Secondary structures have a key role in the structural biochemistry and structural bioinformatics. They are widely used to analyze the protein structures and understand structure assembly process. Two local repetitive structures, i.e., the alpha-helix and the beta-sheet, constitute the vast majority of these secondary structures, while a third infrequent repetitive structure, the turn, had been also been identified and used. The rest is associated to coil. The three repetitive secondary structures have interesting conserved biophysical and geometric properties, while the loops are considered to have no recurrent conformations.

A fourth repetitive structure has also been located even before the turns, the PolyProline II (PPII) helix identified in rich proline sequences. As no simple stabilizing interactions have been found within PPII and its frequency was supposed low, it was considered that this conformation was not so important. However, recent studies shows that PPII frequency is higher than expected, and they could be important for protein – protein interactions. Surprisingly theses new analyses show that PPII are often not composed of Proline.

However, a strong limitation in studding PPII is the absence of consensus assignment. Therefore the aim of this project is to propose a database and webserver dedicated to analyse and assign PPII.

References :

- http://en.wikipedia.org/wiki/Polyproline_helix
- Venkatachalam CM (1968) Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. Biopolymers 6: 1425-1436.
- Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34: 167-339.
- Rose GD (1978) Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 272: 586-590.
- Makowska J, Rodziewicz-Motowidlo S, Baginska K, Vila JA, Liwo A, et al. (2006) Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. Proc Natl Acad Sci U S A 103: 1744-1749.
- Pauling L, Corey RB (1951) The structure of fibrous proteins of the collagen-gelatin group. Proc Natl Acad Sci U S A 37: 272-281.
- Cowan PM, McGavin S, North AC (1955) The polypeptide chain configuration of collagen. Nature 176: 1062-1064.

Présentation des informations à stocker dans le système

La première partie du projet va consister à produire les données à insérer dans la base. L'idée est de traiter les structures de protéines issues de la Protein Databank avec un outil qui nous permettra d'identifier les PPII. Traiter la totalité de la base de données PDB nécessiterait un temps de calcul trop important, aussi nous nous focaliserons sur un sous-ensemble de la PDB. Pour cela, nous allons utiliser l'outil PICSES développé par le laboratoire du Pr. Roland Dunbrack (http://dunbrack.fccc.edu/PISCES.php). PISCES nous permettra d'obtenir une liste de PDB représentatifs à très haute résolution (1.6A) et non redondant (mois de 20% d'identité de séquence) nécessaire pour le travail à réaliser. Si le site PICSES n'est pas accessible, vous utiliserez la liste cullpdb_pc90_res1.8_R0.25_d120106_chains8017 qui est présente dans l'archive (disponible ci-après). Au final, vous vous focaliserez sur les 100 premières structures faisant moins de 1.6A de résolution (vous pourrez également augmenter le nombre de structure, si vous avez le temps).

Par la suite, les fichiers automatiquement téléchargés seront analysés par les outils dssppii.pl, segno, PROSS.py et XTLSSTR-linux contenu dans l'archive accessible à cette adresse :
http://www.dsimb.inserm.fr/~gelly/doc/archive_projet_bdd.tar.gz

Les logiciels annexes nécessaires et les programmes d'extraction sous un format plus simple sont également dans l'archive. Il n'y a pas de fichier de programme d'extraction pour dssppii.pl. Vous pourrez utiliser le programme d'extraction que vos collègues ont développé pour le projet PERL « Prédiction de la structure secondaire ». Vous devrez adapter et modifier certains chemins dans les bibliothèques utilisées par les programmes. Pour vous aider, un README a été préparé dans le répertoire. Les sorties des programmes vous permettront de récupérer les attributions de structures secondaires et plus particulièrement des polyprolines de type II (codé P). Il se peut que certains PDB produisent des erreurs avec les programmes. Dans ce cas, si le nombre de ces PDB n'est pas trop important (moins de 10% de l'ensemble initial), vous pouvez les éliminer sans rechercher les erreurs mais vous préciserez ces PDB pour lesquels une erreur s'est produite. Au minimum, nous demandons pour ce projet au moins deux méthodes fonctionnelles parmi DSSP, SEGNO, PROSS et XTLSSTR (à vous de sélectionner les méthodes qui vous semblent les plus simples à adapter).

La liste des informations utiles et à intégrer dans la base sont les suivantes :

- *PDB id*
- *chain*
- *PDB Header*
- *Amino acid sequence and size*
- *Secondary structure assignment (with PPII of course) by DSSP, SEGNO, XTLSSR*
- *Phi and Psi angle*
- *Resolution*
- …

Cette liste est bien sûr non-exhaustive, vous pouvez rajouter toute autre information que vous jugerez pertinente.

Résumé du projet 💬

Les différentes étapes du projet sont précisées ci-dessous :

Etape 1 : Modélisation du système (ER, relationnel) : vous devrez prendre le temps de réaliser le dictionnaire des données et de créer le modèle conceptuel le plus optimisé possible pour faciliter les futures mises à jour et l'extraction des données à afficher.

Etape 2 : Génération et Extraction des données : Parmi les quatre méthodes citées précédemment, vous devrez au minimum générer les données requises pour deux d'entre-elles (réalisez toutes les méthodes si vous avez le temps). Vraisemblablement, chaque outil fournit la structure secondaire détaillée en exposant notamment les hélices poliprolines (PPII) à partir du fichier PDB.

Etape 3 : Intégration dans la base de données : Vous devrez à partir des programmes utilisés formater les données pour pouvoir les insérer dans la base de données (commande INSERT INTO…).

Etape 4 : Conception de la page d'index puis de l'interface d'interrogation : Votre objectif sera de construire une interface avec divers formulaires permettant d'interroger la base de données (explications détaillées ci-après).

Etape 5 : Conception des sorties : Suite aux formulaires, vous devrez construire une mise en page permettant l'exploitation des données, notamment la comparaison des structures secondaires prédites pour chaque logiciel.

Bonus : Etape 6 : Conception et intégration de sorties complexes : Vous devrez permettre à l'utilisateur d'afficher les cartes de Ramachandran d'un fichier PDB particulier, et ce à partir des coordonnées Phi et Psi contenues dans le fichier PDB (à vous d'inclure ces informations dans la base de données). Pour ce faire, vous pourrez utiliser l'outil MatPlotLib en R (explications détaillées ci-après).

Les programmes additionnels (extraction et « *parsage* » des fichiers de sortie, création de la base de données, insertions automatisées dans la base des informations, génération de graphiques) devront être disponibles dans un répertoire séparé. Egalement, vous détaillerez les commandes SQL permettant la création des tables et les insertions des données dans la base.
L'architecture du `public_html` sera donc la suivante : à la racine, le fichier `index.html` et trois répertoires : `cgi-bin` (pour les programmes cgi), `bin` (pour les programmes non cgi) et `tmp` (pour les données temporaires) et enfin éventuellement `javascript` (pour les javascript).

Conception de l'index

L'index sera une page html classique présentant en anglais l'objectif de la base de données, une introduction aux structures secondaire régulières PolyProline de type II, des exemples d'utilisation du système et des références bibliographiques. Un menu visible sur toutes les pages permettra d'accéder aux différentes pages: Home, Search, Analyse and About. La page about, générée dynamiquement donnera des informations statistiques sur votre base de données (nombre d'acide aminés en structure Polyproline disponibles, nombre de protéines, nombre de requêtes réalisées, par exemple).

1/ Conception de la base d'interrogation

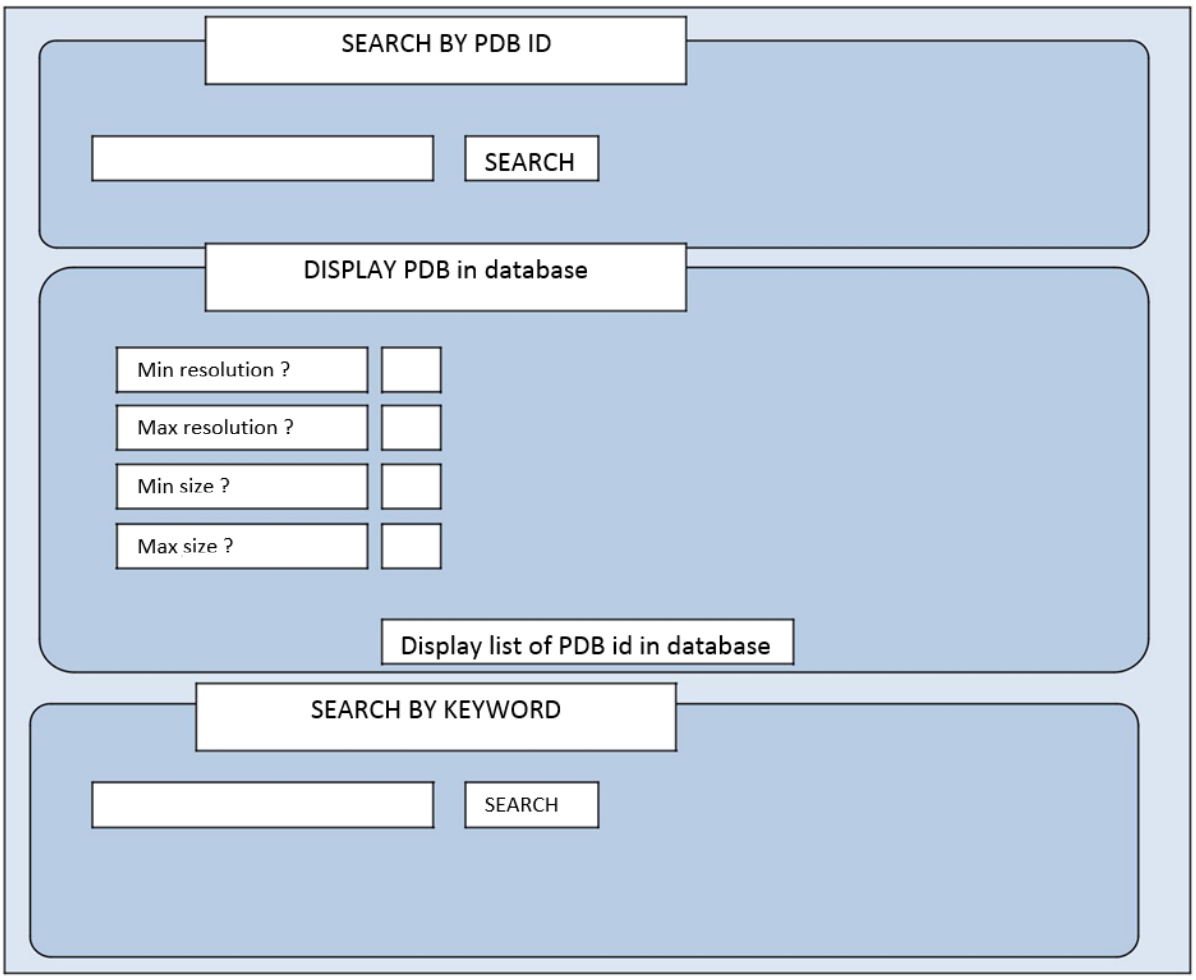


Figure 1 – Formulaires d'interrogation. Le premier doit permettre d'accéder aux résultats d'un PDB ID particulier (ou une liste de PDB). La seconde doit permettre de retourner tous les résultats pour lesquelles les résolutions et tailles minimale et maximale répondent aux critères formulés par l'utilisateur (une page de résultat avec l'ensemble des structures, le pourcentage de PPII, la taille de la protéine et la résolution devra être réalisée). Enfin, le dernier formulaire doit permettre à partir d'un mot-clé de retrouver toutes les structures associées (la recherche doit se baser sur la description de la protéine, un attribut également à inclure dans votre base de données).

Partie 1 : Cette interface devra permettre à un utilisateur externe se connectant via un navigateur, d'interroger le système selon un PDB ID ou plusieurs PDB ID séparés par un espace ou un retour à la ligne. Attention à la sécurité : chaque donnée entrée dans le formulaire devra

être soigneusement analysée. Possiblement, un tableau dynamique permettra de trier les PDB retournés (dans le cas de plusieurs PDB retournés) selon un ordre alphabétique des identifiants, la taille de la protéine, etc. (Javascript -> Bibliothèque Jquery). Si un seul résultat n'est retrouvé, alors on affiche directement la page de résultat détaillée relative à ce PDB (voir 3/ ci-après). **Figure 1 – formulaire 1**.

Partie 2 : Une liste des PDB présents dans la base pourra être affichée en cliquant sur le bouton dédié, à partir des résolutions et tailles minimales/maximales. Cette liste sera générée dynamiquement en interrogeant la base. Les résultats devront s'afficher de façon similaire à la partie 1 (avec un tableau dynamique, si possible). **Figure 1 – formulaire 2**.

Partie 3 : La troisième partie de l'interface devra permettre d'interroger la base par mots clés dans le *header*. Les résultats affichés devront adopter un style similaire à celui de la partie 2 (avec un tableau dynamique, si possible). **Figure 1 – formulaire 3**.

2/ Conception de la page de résultat lors d'une requête issue de l'interrogation

Après exécution de la recherche, une page de résultats s'affichera, générée de manière dynamique.
S'il n'y a pas de PDB ID correspondant à la requête, un message adapté devra être affiché.
La page de résultat générée affichera une liste qui comprendra les PDB ID avec leur header, la résolution et la taille ainsi que les liens vers PDBsum (**Figure 2**). Un lien spécifique lancera l'interrogation de façon automatique et détaillée à l'aide d'un lien avec le PDB ID concerné (voir 3/ ci-après). Possiblement, vous utiliserez une méthode basée sur javascript (jquery/tablesorter) pour choisir l'ordre du tri de chaque colonne. Si un seul résultat n'est trouvé, alors on passera directement à l'affichage des résultats détaillés, il est inutile de présenter ce tableau (voir 3/ ci-après).

Search:

PDB	Chain	Title	Length	Resolution	PPII (n)	PPII (%)
1A4S	A	OXIDOREDUCTASE	503	2.10	20	4.0
1ATI	A	PROTEIN BIOSYNTHESIS	505	2.75	15	3.0
1BBU	A	LIGASE	504	2.70	18	3.6
1BU6	O	TRANSFERASE	501	2.37	4	0.8
1BXS	A	OXIDOREDUCTASE	501	2.35	17	3.4
1DL2	A	HYDROLASE	511	1.54	16	3.1
1DPE	A	PEPTIDE TRANSPORT	507	2.00	20	3.9
1E1O	A	LIGASE	504	2.12	19	3.8
1E4M	M	HYDROLASE	501	1.20	10	2.0
1ECF	A	TRANSFERASE (GLUTAMINE AMIDOTRANSFERASE)	504	2.00	0	0.0
1EI5	A	HYDROLASE	520	1.90	19	3.7
1EZ0	A	OXIDOREDUCTASE	510	2.10	20	3.9
1GCY	A	HYDROLASE	527	1.60	9	1.7
1GKM	A	LYASE	507	1.00	14	2.8
1GPM	A	TRANSFERASE (GLUTAMINE AMIDOTRANSFERASE)	525	2.20	18	3.4
1GWE	A	OXIDOREDUCTASE	503	0.88	27	5.4
1GYT	A	HYDROLASE	503	2.50	9	1.8
1HCU	A	GLYCOSYLATION	503	2.37	15	3.0
PDB	Chain	Title	Length	Resolution	PPII (n)	PPII (%)

Showing 1 to 50 of 350 entries

Figure 2 – Page de résultats suivant une requête à partir des différents formulaires (il est évidemment possible d'obtenir un unique résultat). En cliquant sur le code PDB, on accédera aux résultats détaillés relatifs au PDB.

3/ Conception de la page de résultat pour un PDB détaillé

Lorsque vous accédez à un résultat détaillé, vous devez tout d'abord exposer les résultats des prédictions de chaque méthode à travers un alignement (**Figure 3**). Egalement, vous devez indiquer un lien vers PDBsum en accord avec la structure de la protéine étudiée. Un code couleur en accord avec les structures secondaires sera utilisé (type Helix rouge, type Bêta vert, type Coil gris et type PPII en bleu et souligné) pour surligner les différents types de structures secondaires mises en évidence par l'attribution (vous pouvez bien sûr choisir votre propre code couleur).

Figure 3 – Page de résultat détaillé pour un PDB ID précis. Dans un premier temps, on affiche l'alignement des différentes prédictions de chaque méthode (deux au minimum). La première ligne correspond à la séquence de la protéine. Un code couleur est associé pour chaque structure secondaire (hélice alpha, brins beta, coil, polyproline). En dessous, un graphe de *Ramachandran* pour chaque méthode réalisé devra être affiché (à l'aide de MatPlotLib, mais libre à vous de trouver une alternative en Python si vous le souhaitez). Evidemment, vous pouvez simplement construire un simple graphe de -180 / +180 en mettant en évidence les régions avec une structuration secondaire particulière.

Un graphe de *Ramachandran* sera généré pour chaque méthode en suivant le code couleur utilisé précédemment. Pour cela, un répertoire de nom aléatoire est créé dans le repertoire `tmp` (en utilisant la fonction random : ex ppii.1242342). Ce répertoire servira à stocker les données temporaires tel que les scripts R et les graphiques png créés. Ce répertoire sera régulièrement effacé (tous les jours). Une autre solution serait de réaliser un simple graphe (sans image à générer) en mettant en évidence les régions sous une structuration particulière. S'agissant du graphe de *Ramachandran* proprement dit, il fera correspondre à chaque acide aminé un point de couleur ou le numéro de l'acide aminé entouré d'un cercle (plus difficile). Les Polyprolines de type II devront être particulièrement mis en valeur.

Pour aller plus loin…

1/ Analyse directe depuis un PDB :

Si vous souhaitez améliorer votre application, vous pourrez créer un nouveau formulaire avec lequel vous pourrez soumettre directement un fichier PDB. Si le code PDB associé est déjà présent dans la base de données, les résultats seront dès lors renvoyés sous forme détaillée. Dans le cas contraire, il faudra permettre le *parsage* et l'extraction des données, pour ensuite les ajouter à la base. Le tout doit pouvoir se faire automatiquement et renvoyer la page de résultat détaillée. Vous pourrez utiliser le répertoire *tmp* pour stocker le fichier PDB, lequel sera ensuite supprimé au bout d'une journée.

2/ Interface privée :

Concevez une interface protégée par login et mot de passe, permettant de lancer directement des requêtes SQL. Ajoutez un lien à partir du menu et de l'index.
Cette interface permettra également d'insérer de nouvelles données à partir d'un fichier de sortie. Une vérification des données devra être réalisée pour ne pas insérer de données déjà présentes ou des données erronées (CHAR à la place d'un INT par exemple).

3/ Interface dynamique :

Modifier l'interface d'interrogation du formulaire 1 afin de permettre au fur et à mesure de suggérer et d'afficher une liste des PDB ID dont les lettres commencent par les première lettres saisies (d'une manière identique à *Google suggest*). Il est probable que vous ayez à utiliser des méthodes se basant sur du javascript (*jquery* par exemple).

Rapport :

Outre le site, nous vous demandons la rédaction d'un court rapport (environ cinq pages) dans lequel vous exposerez les outils utilisés, jusqu'où vous avez pu aller et ce qui ne fonctionne pas (en effet, même si votre application ne fonctionne pas, nous aimerions comprendre où cela a pu coincer). Le rapport devra suivre un fil similaire à une revue scientifique : une introduction qui rappelle le contexte et l'objectif de l'application ; une partie matériels/méthodes qui présentera le MCD de votre base de données, les divers langages utilisés et dans quels objectifs (Python, R pour les graphes de Ramachandran, etc.) ; la partie Résultats devra exposer quelques captures d'écran de votre application ; et la discussion/conclusion servira à indiquer les limites de votre projet et ce que vous pourriez tenter d'améliorer à l'avenir.

Besoin d'aide ?

En cas de soucis, vous pouvez bien évidemment nous envoyer un mail, aussi bien pour des questions concernant l'installation de MySQLdb (qui pose souvent des problèmes), sur l'extraction des données ou même sur la conception de l'application.
romain.coppee@hotmail.fr (ou romain.coppee@univ-paris-diderot.fr)
costas.bouyioukos@univ-paris-diderot.fr