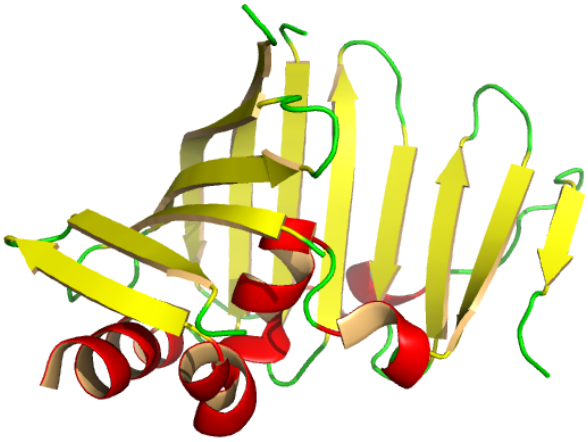
Protein Peeling



Projet Long

Rapport

Sonia Tieo
M2BI

université PARIS DIDEROT PARIS 7

7 janvier 2019

Table des matières

1 Intoduction

Les protéines sont des macromolécules biologiques du vivant essentielles pour la réalisation de diverses fonctions. Elles peuvent être décrites à partir de quatre organisations structurales.
La structure primaire est définie comme des polymères linéaires d'acides aminés selon Fischer et Hofmeister (1902). Cette chaîne peut se déformer (torsion ou repliement) donnant lieu à une structure secondaire'(hélices ou feuillet). Certaines régions sont non structurées (boucles, tours, coudes...).
Certaines protéines forment une structure tertiaire (enchaînement de structure secondaire) voire une structure quaternaire (enchaînement de structure tertiaires) lui conférant des propriétés et fonctions physiologiques.
Ainsi, les protéines, tout d'abord défini comme des séquences peuvent déterminer la structure tridimensionnelle d'après les recherches d'Anfinsen.

Actuellement, les structures protéiques sont principalement résolues par cristallographie aux rayons X ou par la résonance magnétique nucléaire (RMN). La microscopie électronique peut aussi être utilisé. Toutes les structures sont recensés dans une base de donnée : PDB (Protein Data Bank). [1] Connaitre la structure 3D est donc importante et permet de comprendre par exemple leur rôle dans certaines maladies ou encore pour la recherche de médicaments efficaces.[2]

L'organisation des structures protéiques 3D peut être représenté comme un assemblage de différentes structures secondaires.[3]. En effet, ces structures forment des motifs particuliers que l'on retrouve dans de nombreux repliements protéiques, ce sont les structures super secondaires. Nombre d'entre elles ont été bien caractérisées, telles que les épingles à cheveux β [4], les enchaînements hélice-boucle-hélice ou $feuillet\beta–helice\alpha–feuillet\beta$, lesfaisceaux de quatre hélices (souvent α) ou des tonneaux β comme par exemple le mode de repliement dit en « clé grecque » [5]. (voir Figure 1)
Une autre manière de voir les structures protéiques est sa décomposition en domaines fonctionnels conservés.

De nombreux auteurs ont proposé différentes méthodes pour scinder les structures protéiques en domaines. Par exemple, l'outil PDP : protein domain parser [6] permet l'identification automatique de domaines selon trois critères différents : (i) par comparaison avec la base de données SCOP organisée par des experts des domaines ; (ii) par comparaison avec une collection d'assignations manuelles de domaine et (iii) par comparaison avec un ensemble de 55 protéines, fréquemment utilisé comme référence pour l'attribution automatique de domaine.

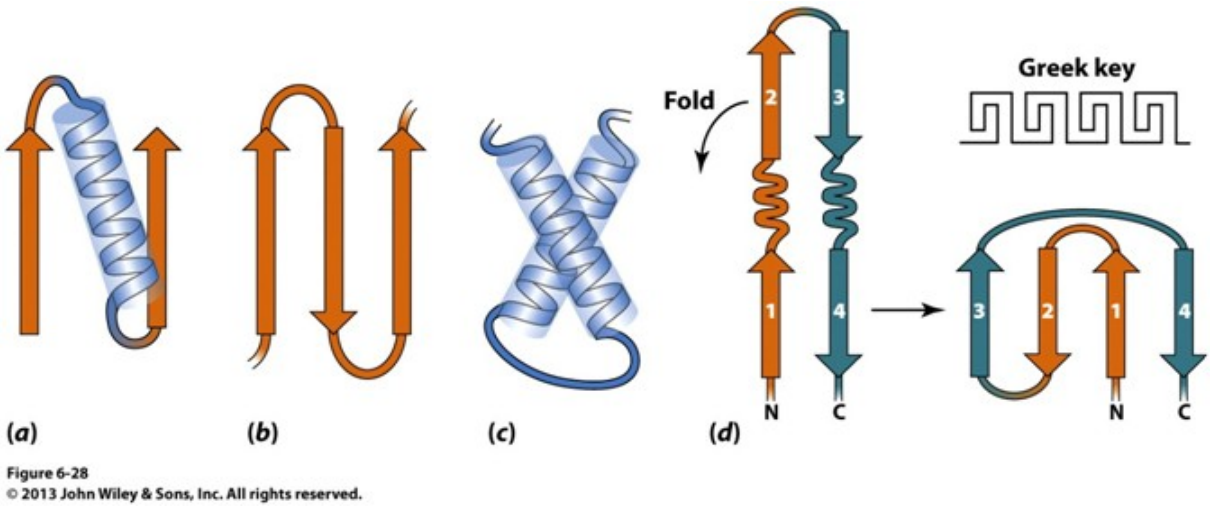


Figure 6-28

FIGURE 1 – Structure super secondaire.
(a) hélice-boucle-hélice, (b) épingles à cheveux β, (c) hélice-boucle-hélice, (d) clé grecque

Cependant les domaines sont souvent de grandes tailles donc des auteurs se sont penchés sur des outils permettant de découper hiérarchiquement les protéines en plus petites unités.
Les auteurs de DIAL [7] commencent par regrouper les structures secondaires (hélices et feuillets) en fonction de la distance des carbones α des résidus traduit par un "indice de proximité". L'organisation des structures secondaire est représenté sous la forme de dendrogrammes où les noeud identifiés sont des clusters de structures tertiaires (incluant domaines et structure super-secondaires).
Quant aux auteurs de Protein Peeling [8], ils cherchent à découper la structure 3D en plusieurs unités protéiques (UP). Ce niveau intermédiaire d'organisation, entre les structures secondaires et les domaines, est défini comme une sous-région compacte et indépendante. Le principe de base de Protein Peeling est d'obtenir des UPs avec un grand nombre de contacts intra-UP et un faible nombre de contacts inter-UP. Pour les identifier, cette technique réalise une série de partitions imbriquées successives. Cela conduit à la construction d'une arborescence (ou d'une hiérarchie) montrant la division successive des unités en sous-unités.

L'objectif du projet est de proposer une nouvelle méthode de Protein Peeling sans utiliser une segmentation hiérarchique mais en recherchant itérativement les régions de la protéines les plus compactes et les plus indépendantes.

Nous expliciterons l'implémentation de la méthode dans un premier temps puis nous vérifierons la cohérence sur des exemples que nous comparons avec le programme Protein Peeling d'origine. (accès : http ://www.dsimb.inserm.fr/dsimb_tools/peeling3/)

2 Matériel et Méthode

L'outil a été été codé en Python 3 et se lance en ligne de commande bash. Il nécessite en argument un fichier PDB (Protein Data Bank) contenant les informations de la structure 3D d'une protéine. L'utilisateur devra préciser la chaîne sur laquelle il souhaite travailler ainsi que la taille minimale et maximale d'un PU. (voir précision dans la partie : Instruction pour utiliser le programme)

2.1 Matrice de probabilité de contact

Tout d'abord, la lecture et le traitement du fichier PDB a été effectué à l'aide du module BioPython (avec Bio.PDB).
Une matrice de distance euclidienne des carbones α des résidus pris deux à deux est calculée à partir des coordonnées atomiques cartésiennes.
A partir de cette matrice, une matrice de probabilité de contact (voir Figure 2) est implémentée d'après la transformation logistique suivante :

$$\boxed{p(i,j) = \frac{1}{1 + exp(\frac{d(i,j) - d_0}{\Delta})}}$$

Chaque distance $d(i,j)$ entre deux résidus i et j est transformée en probabilité $p(i,j)$ avec d_0 et Δ valant respectivement 8 et 1,5 Å.
Ainsi, plus une distance est faible, plus la probabilité de contact sera proche de 1.

2.2 Découpage optimal de la protéine en unités protéiques (UP) en fonction de l'index de partitionnement (PI)

On rappelle qu'une UP est un fragment protéique compact et indépendant, ie : où les contacts entre les résidus au sein d'une UP sont nombreux comparés aux contacts de ces résidus avec le reste de la protéine .
La recherche d'UP est réalisée itérativement et certains critères doivent être remplis :
- Une UP est définie entre deux positions [a,b]. (Figure 2)
- Une UP ne doit pas couper une structure secondaire (hélices et feuillets)
- Une UP possède une taille minimale et maximale (choisis pas l'utilisateur)

Pour définir les structure secondaires, DSSP [9] a été utilisé. L'outil utilise la version implémenté par Biopython.
Ici nous considérons que les structures non secondaires regroupent les ponts β , les boucles et les tours ainsi que les "coil".

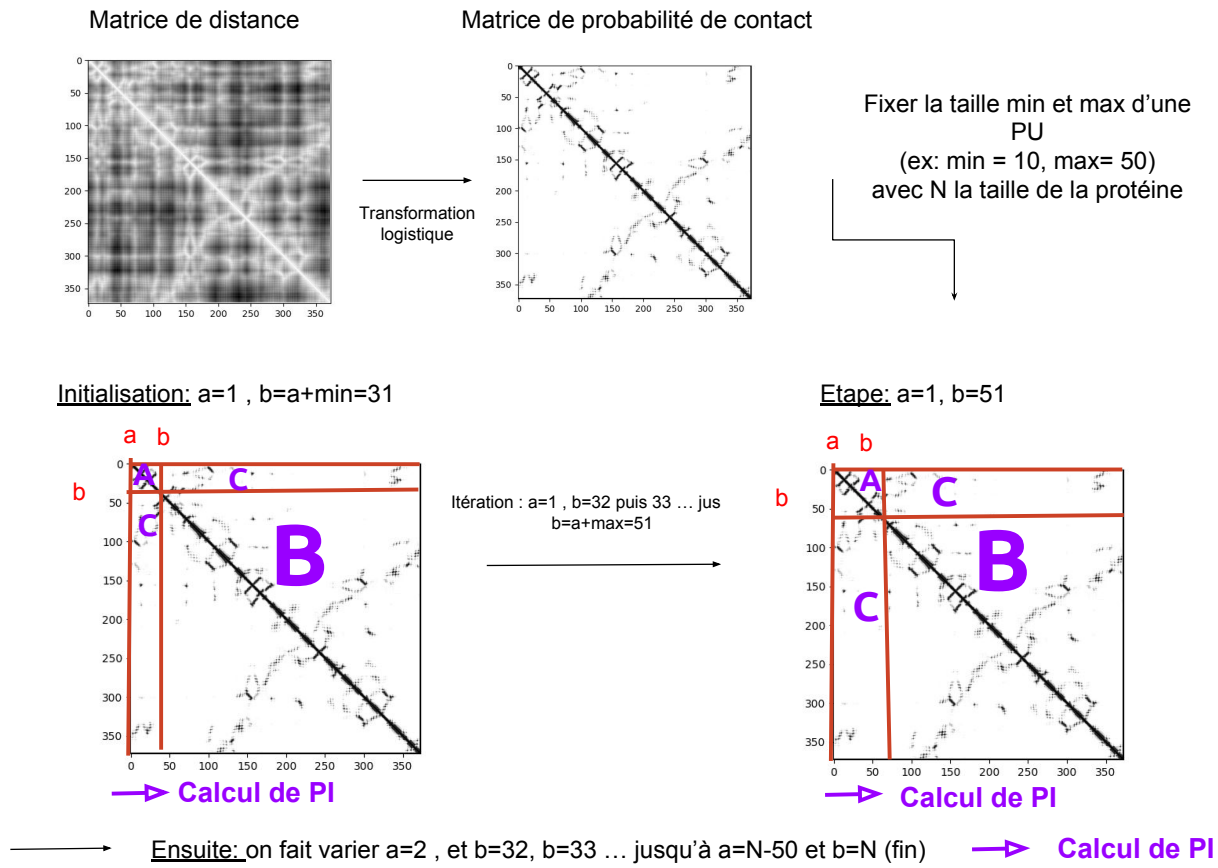


FIGURE 2 – Pipeline du programme - Première partie.
Construction de la matrice de distance et de probabilité de contact puis évaluation des PI pour des UPs potentiels.

Soit une protéine de taille **N**.
Initialement, la première position **a** est fixé à 1 : a=1 et la position **b** varie dans l'intervalle : [a+min,a+max]. (Figure 2)

Pour chaque couple de valeurs **a** et **b** est calculé une valeur de PI défini par :

$$PI_{a,b} = \frac{AB - C^2}{(A + C)(B + C)}$$

Avec, **A** la somme des probabilités de la sous-matrice carrée aux dimensions [a :b,a :b], **B** la somme des probabilités de la sous-matrice carrée aux dimensions [b :N,b :N] et **C** la somme des probabilités de la sous-matrice aux dimensions[b :N,a :b].

Cette valeur de PI est une valeur sous-jacente au coefficient de corrélation de Matthews (MCC). Elle permet d'évaluer ici la présence de nombreux contacts au sein des

sous-unités et un nombre limité de contacts entre sous-unités.

Ensuite la position **a** varie : a=2 et on refait varier **b** entre [a+min,a+50] pour calculer le PI.
Le processus est itéré pour **a** variant de [1,N-max] tout en faisant varier **b** pour chaque valeur de **a**.

Ainsi, pour chaque position de **a** donnée correspond une liste de valeur de PI de taille **max**, en fonction de chaque valeur de **b** associée.

Sur cette liste de valeurs est récupérée la valeur maximale de PI. Ainsi pour chaque résidu de départ**a** correspond un résidu de fin **b** qui forme un UP qui maximise la valeur PI.

2.3 Obtention de tous les découpages possibles

La suite du programme va "assembler" toutes les combinaisons d'UP possibles.
En principe, tous les résidus de départ potentiels **a** sont parcourus. Pour chaque **a** correspond un résidu de fin **b** . On regarde si **b** peut initier une UP (ie : **b** devient le résidu de "départ" et on regarde s'il est associé à un résidu de fin). On le fait jusqu'à ce que ce ne soit plus possible.

Après obtention de toutes les combinaisons possibles, on éliminera celles qui ne respectent pas les critères pour être un UP. (voir Figure 3)

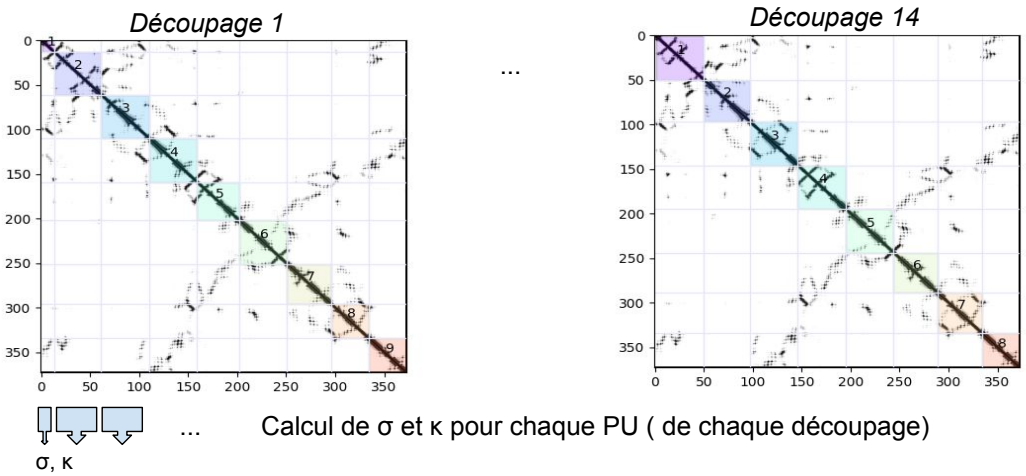
2.4 Sélection du meilleur découpage en fonction des critères de compacité et de séparation

Pour discriminer les découpages, deux critères supplémentaires sont implémentés : la séparation σ et la compacité κ. [10]. (Figure 3)

Le critère de séparation σ mesure l'indépendance entre 2 UPs : UP-A (définis entre les résidus a et b) et UP-B (définis entre les résidus b et c). Elle est exprimée d'après par la formule suivante avec A, B, C la somme de toutes les probabilités de contacts au sein de chaque sous matrice (schématisée sur la Figure 4) :

$$\sigma = \frac{\frac{C^2}{taille_A * taille_B}}{\frac{A+B+C}{taille_A * taille_B * taille_C}}$$

1) Couples (a,b) qui maximise le PI d'un potentiel PU entre a et b
2) "Assemblage" des potentiels PUs pour constituer la protéine.
3) Différents découpages:



4) Sélection du découpage en fonction du meilleur critère σ et κ global (somme des σ et κ de chaque PU)
5) Représentation graphique des PUs du meilleur découpage (couleur pas PU) avec Pymol

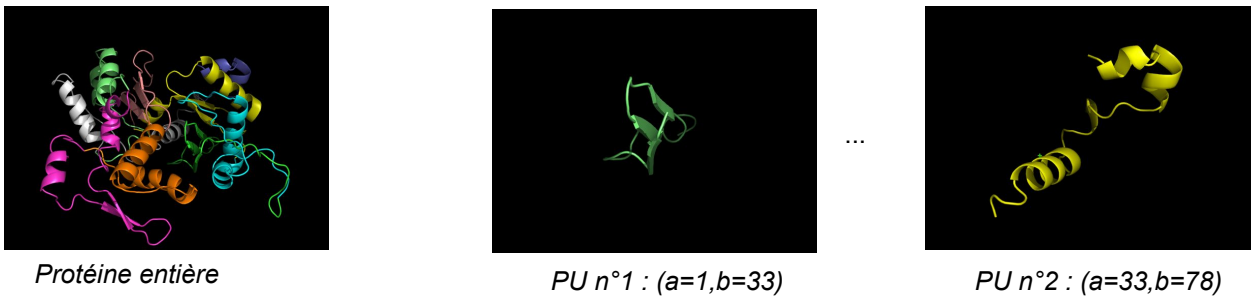


FIGURE 3 – Pipeline du programme - Deuxième partie.
Construction de la matrice de distance et de probabilité de contact puis évaluation des PI pour des UPs potentiels.

Étant donné qu'une grande valeur de σ représente un nombres de contact élevé entre les UPs, on voudra un faible σ. Ainsi, plus la valeur de σ est faible plus les unités protéiques UP-A et UP-B sont indépendants.

Quant au critère κ, il calcule la compacité. Il mesure la densité des contacts au sein d'une UP (ex : Pu-A). Il est exprimé comme suit avec A la somme des probabilités dans la sous matrice :

$$\kappa = \frac{A}{taille_A}$$

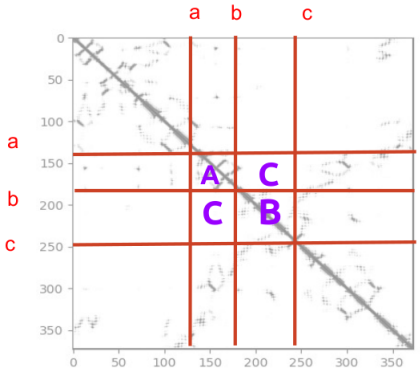


FIGURE 4 – Découpage centrée sur 2 UPs : UP-A[a,b] et UP-B[b,c] pour illustrer le calcul des critères σ et κ

Ainsi, plus la valeur de κ est grande, plus l'UP est compacte.

Au final pour sélectionner le meilleur découpage, une moyennes de valeurs de kappa d'une part et des valeurs de σ d'autre part sont calculées. A partir de ces moyennes, les découpages sont ordonées et on sélectionne celui qui possède la moyenne minimale de σ et la moyenne maximale de κ.

2.5 Instruction pour utiliser le programme

Dans le répertoire **Protein_Peeling**, se trouve le fichier **ProteinPeeling.py**. Il est recommandé de placer le fichier PDB sur lequel vous voulez travailler dans le sous-répertoire **datas**. Pour executer (à la racine), veuillez entrer :

```
python3 Protein_Peeling.py PDBfile CHAIN MIN MAX
```

Avec :
- PDBfile : Votre fichier PDB
- CHAINE : La lettre de la chaîne sur laquelle vous souhaitez travailler
- MIN : la taille minimale d'une UP
- MAX : la taille maximale d'une UP

Par exemple :

```
python3 Protein_Peeling.py /datas/1atn.pdb A 10 50
```

L'outil génère dans **results** un répertoire nommé en fonction du code PDB. Ce répertoire contient la matrice de contact avec le meilleur découpage et un fichier **ranking_spitting.txt** qui classe et donne les informations (σ , κ moyens ...) sur les découpages proposés. Chaque découpage est repertorié dans un dossier : PU1, PU2 qui

correspond au découpage 1, au découpage 2 ...
Dans le dossier PU1, par exemple, se trouvent les PDBs divisés et les figures des structures 3D de la protéine entière et des UPs ainsi qu'une courbe des crières σ , κ et PI.

Pour des informations supplémentaires, un **README** et une documentation **ProteinPeeling.html** sont fournis avec l'outil.

3 Résultats

3.1 Application sur une petite protéine : ubiquitine (1ubi)

3.1.1 Comparaison du meilleur découpage et du moins

Nous choisissons l'ubiquitine humaine (de code PDB : 1ubi) qui est une protéine de petite taille : 76 acides aminés. [11]

Après exécution du programme sur 1ubi en choisissant une taille de PU de 6 résidus minimum et de 20 réisidus maximum, nous obtenons les découpages résumés dans le Tableau 1.

PU	Rang	Decoupage	Sigma	Kappa
1	1	[1, 21, 50, 77]	1.1267 ± 0.49	7.8837 ± 0.6865
2	8	[1, 10, 37, 61, 77]	1.3489 ± 0.1591	6.5648 ± 1.3221
3	7	[1, 11, 37, 61, 77]	1.2858 ± 0.1695	6.7351 ±
4	3	[1, 17, 46, 77]	1.2855 ± 0.6045	7.9063 ± 1.1375 0.4362
5	5	[1, 18, 47, 77]	1.2814 ± 0.5872	7.854 ± 0.4444
6	6	[1, 19, 47, 77]	1.2504 ± 0.5613	7.8413 ± 0.5027
7	4	[1, 20, 47, 77]	1.2228 ± 0.5538	7.8482 ± 0.5648
8	4	[1, 21, 50, 77]	1.1267 ± 0.49	7.8837 ± 0.6865

TABLE 1 – Combinaisons de découpage de 1ubi pour les paramètres : taille UP min = 6 et taille UP max = 20. Le rang, la combinaison des UPs et les valeurs moyenne de κ et σ sont renseignés.

Nous pouvons comparer deux découpages possibles :
- Découpage n°6 (Rang 1 : meilleur découpage) (voir Table 1)
- Découpage n°2 (Rang 8 : moins bon découpage) (voir Table 1)

Le découpage n°6 comporte 5 UPs et le découpage n°2 comporte 6 UPs. Quantitativement et de manière globale, la valeur moyenne de κ pour le découpage 6 est d'environ 6.7625, soit environ 0.8717 de plus que le découpage 2. Quant à la moyenne σ, elle vaut 0.7919 pour le meilleur découpage soit 0.3566 de moins que celle du moins découpage.
Pour affirmer que le découpage 6 est significativement meilleur que le 2ème, un test statistique de comparaison de moyenne pour κ d'une part et pour σ d'autre part est réalisé.

Nous choisissons un z-test pour petit échantillon (moins de 30) car on connaît la variance de nos deux échantillons.

Pour κ, on pose l'hypothèse H0 qui stipule que la moyenne des κ du découpage 6 (κ_6) peut être considérée comme égale à κ_2 (celui du découpage 2). L'hypothèse alternative H1 : $\kappa_6 > \kappa_2$. On retient un risque de première espèce de 10%. (moins stringeant)
Dans ce test, Z=1.3439 et p-value = 0.08949. Comme la p-value est inférieure à 0.1, on rejette H0 avec 10% de chance de se tromper et on conclut que κ_6 est significativement supérieur à κ_2.
En suivant le même raisonnement pour σ, on teste si : $\sigma_6 < \sigma_2$. Cette fois ci, on ne peut pas rejeter H0 ($\sigma_6 = \sigma_2$) au même seuil au de 10%. Donc les valeurs moyennes de σ ne sont pas significativement différentes.

Néanmoins, malgré une différence de moyenne de σ moins convaincante, nous pouvons tout de même dire que le découpage n°6 est meilleur que le découpage n°2. En effet, si nous examinons les valeurs de σ et κ de chaque UP des 2 découpages, nous voyons, que le découpage du début n'est pas optimal. (voir Figure 5)

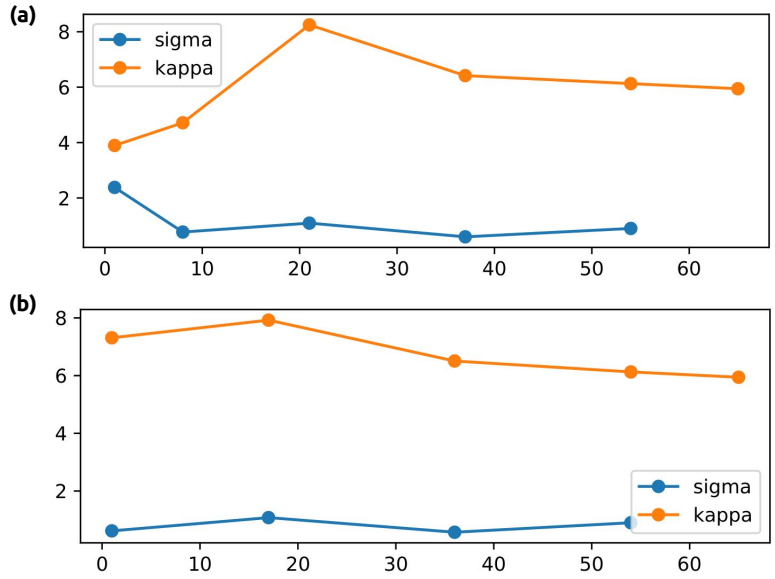


FIGURE 5 – Comparaison des valeurs de σ et κ pour chaque UP (a) du découpage 2 et (b) du découpage 6.
Pour lire la figure : pour (a) Le point orange pour abscisse 1 correspond à la valeur de κ du UPs d'intervalle[1 :11] en ordonnée et le second point orande pour abscisse 11 correspond à la valeur de κ du UPs d'intervalle[12 :21] en ordonnée ...

D'après la Figure 5 (b), le découpage 6 montre des valeurs de κ pour chaque UPs avoisinnant 8 (optimal) au début (entre les résidues 1 à 18) et 6 vers la fin. En comparaison,

le découpage 2 en (a) montre entre les résidus 1 à 21, une valeur de kappa faible, ie : environ 4 pour l'UP de [1,8] et de 4.5 de [9,21].

Ainsi, d'après ces deux cas, pour le découpage de l'ubiquitine, il est préférable de garder une UP plus grande entre les résidus 1 à 21 car elle est plus compacte et plus indépendante (σ en (b) plus faibles qu'en (a)) que deux petites UPs entre 1 à 8 puis 9 à 21.

Ce qu'il manquerait ici comme analyse quantitative est la comparaison des moyennes entre tous les découpages possibles.

3.1.2 Impact de la taille min et max de l'UP.

Nous allons maintenant aborder un autre paramètre : la taille des UPs.

L'utilisateur a la possibilité de définir la taille minimale et maximale du UP. Nous allons comparer l'impact sur l'ubiquitine en examinant deux cas :
- le cas précédent où la taille minimale d'une UP est de 6 et la taille maximale de 20
- le cas où la taille minimale est de 10 et la taille maximale de 30. (résumé sur la Table 2)

Découpage	**Rang**	**Combinaison**	**Moyenne σ**	**Moyenne κ**
1	1	[1, 21, 50, 77]	1.1267 ± 0.49	7.8837 ± 0.6865
2	8	[1, 10, 37, 61, 77]	1.3489 ± 0.1591	6.5648 ± 1.3221
3	7	[1, 11, 37, 61, 77]	1.2858 ± 0.1695	6.7351 ± 1.1375
4	3	[1, 17, 46, 77]	1.2855 ± 0.6045	7.9063 ± 0.4362
5	5	[1, 18, 47, 77]	1.2814 ± 0.5872	7.854 ± 0.4444
6	6	[1, 19, 47, 77]	1.2504 ± 0.5613	7.8413 ± 0.5027
7	4	[1, 20, 47, 77]	1.2228 ± 0.5613	7.8483 ± 0.5648
8	1	[1, 21, 50, 77]	1.1267 ± 0.49	7.88375 ± 0.6865

TABLE 2 – Combinaisons de découpage de 1ubi pour les paramètres : taille UP min = 10 et taille UP max = 30. Le rang, la combinaison des UPs et les valeurs moyenne de κ et σ sont renseignés.

D'un point de vue général, comme la taille des UPs est plus grande, les découpages entraînent des UPs plus grandes. Par exemple, d'après la Table 2, le meilleur découpage, la 1ère , trouve 3 UPs contrairement au meilleur découpage du cas précédent (où les UPs sont plus petites) qui en trouve 5. (visualisation des 2 structures ainsi que de la matrice de contact découpée sur la Figure 6).

De la même manière, un test de comparaison de moyenne des 2 cas peut être réalisé avec un z-test.

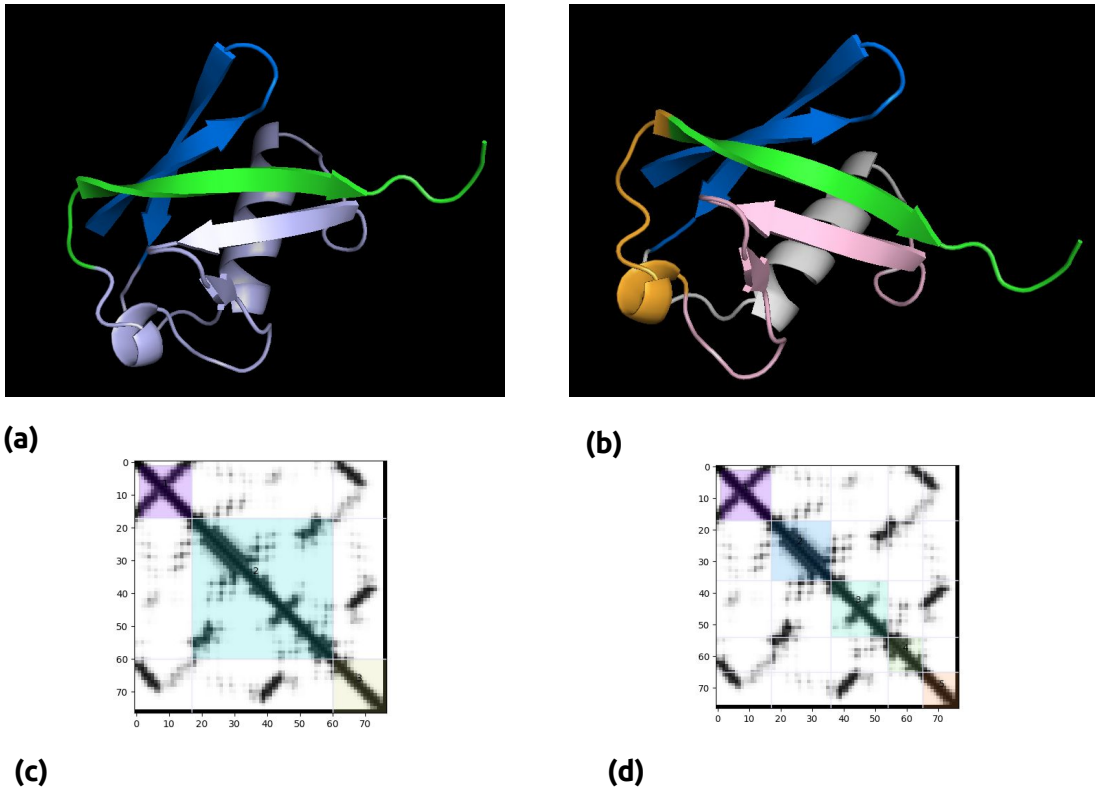


FIGURE 6 – Représentation de 1ubi
(a) Protéine découpée optimalement en 3 UPs pour les paramètres : taille PUmin = 10 et taille PUmax = 30 et sa matrice de probabilité de contact découpé en (c)
(b) Protéine découpée optimalement en 4 UPs pour les paramètres : taille PUmin = 6 et taille PUmax = 20 et sa matrice de probabilité de contact découpé en (d).
Représentation structurale sur Pymol et des matrices sur Python

Soit κ_{petit} la moyenne des κ du meilleur découpage permettant des petites UPs (min=6, max=20) et κ_{grand} celle du meilleur découpage permettant des grandes UPs (min=10, max=30). D'après un z-test pour un risque de 10%, on ne peut pas rejeter l'égalité des moyennes. Donc à priori, les deux découpages sont tous les deux bons si on se réfère aux valeurs de κ.
Pour pouvoir trancher, il faudra une argumentation biologique.

3.2 Application sur une grande protéine : l'actine (chaine A de 1atn)

Le fichier PDB de code 1atn représente un complexe entre la chaine A qui correspond à l'actine de lapin et la chaine D qui correspond Deoxyribonuclease I (DNASE-I) bovin. La structure de ce complexe a été bien étudié.[12] Elle est constitué de 633 résidus et 371 composent l'actine.

Après exécution du programme, nous trouvons que le meilleur découpage est celui en Figure 7 avec 8 UPs avec une valeur de κ moyenne de 9.2138 +/- 0.7893 et une valeur de σ moyenne de 0.7252 +/- 0.3014.

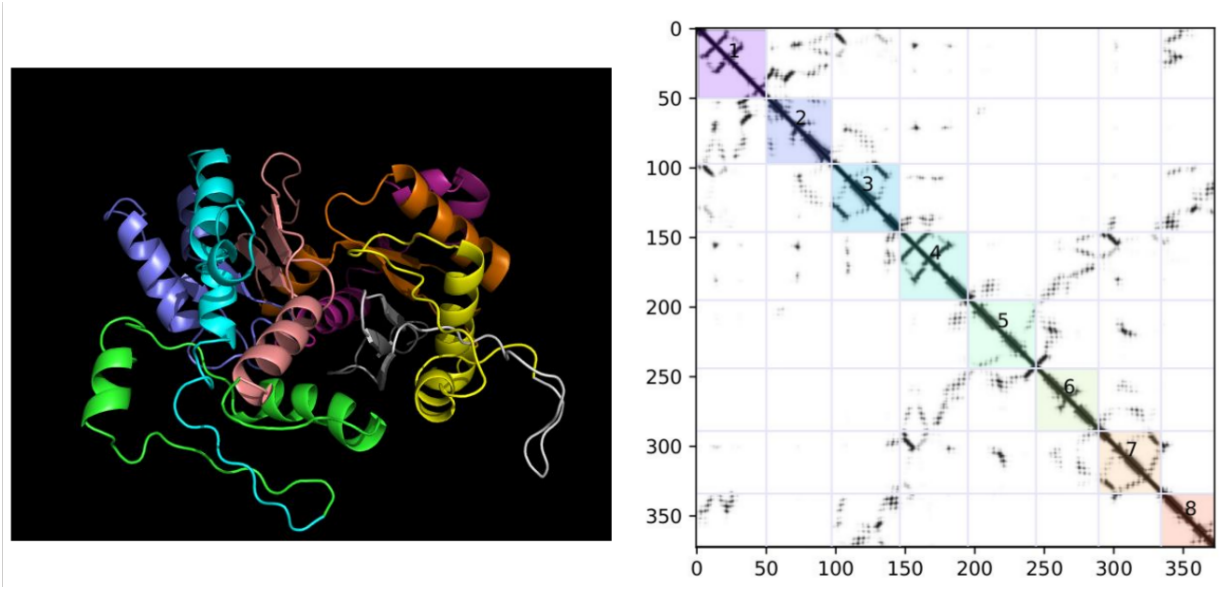


FIGURE 7 – Représentation de 1atn
Protéine découpée optimalement en 8 UPs pour les paramètres : taille PUmin = 10 et taille PUmax = 50 et sa matrice de probabilité de contact découpé
Représentation structurale sur Pymol et des matrices sur Python

On retrouve des motifs remarquables dont certains de structures super secondaires connus comme par exemple le motif "hélice-boucle-hélice" ou encore un motif " feuillet-hélice-feuillet-hélice". (Figure 8)

Nous verrons la pertinence de ces UPs en Discussion.

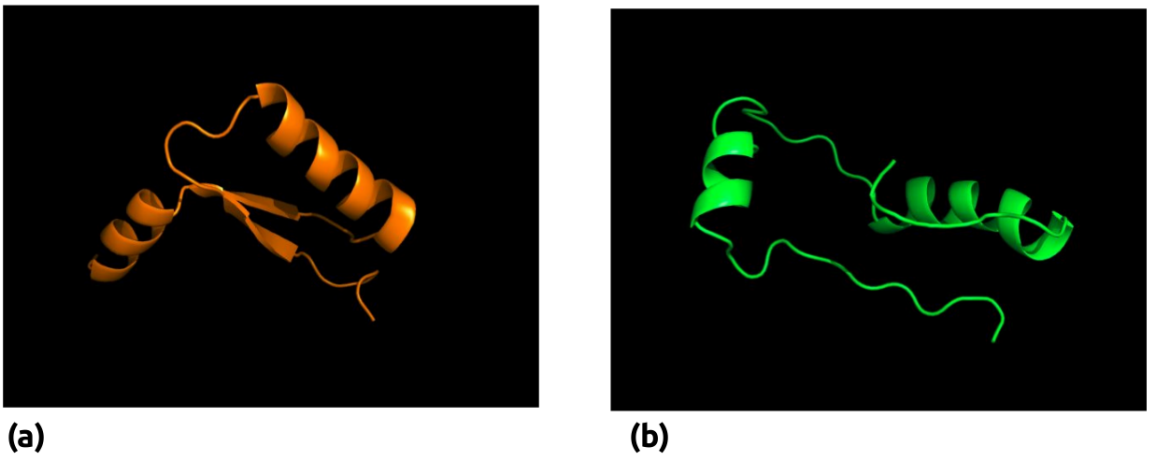


FIGURE 8 – Représentation de 2 UPs de 1atn
(a) UP entre les résidus 97 et 146 , (b) UP entre les résidus 195 et 244
Composition : [1, 50, 97, 146, 195, 244, 289, 334, 373]
Représentation structurale sur Pymol et des matrices sur Python

4 Discussion

4.1 Comparaison du découpage de l'ubiquitine avec le résultat de Protein Peeling et la littérature

Nous rappelons que le nombre de UPs qui constituent l'ubiquitine dépend des paramètres de base fixé pour la taille minimale et maximale des UPs.
Nous n'avons à réussi à trancher entre le découpage en 3 ou 5. (Figure 6).

Les bases de donnée CATH (Protein Structure Classification database) [13] et SCOP (structural classification of proteins database) [14] définissent la protéine entière comme étant un unique domaine. Nous n'avons donc pas d'information sur les domaines de 1ubi.

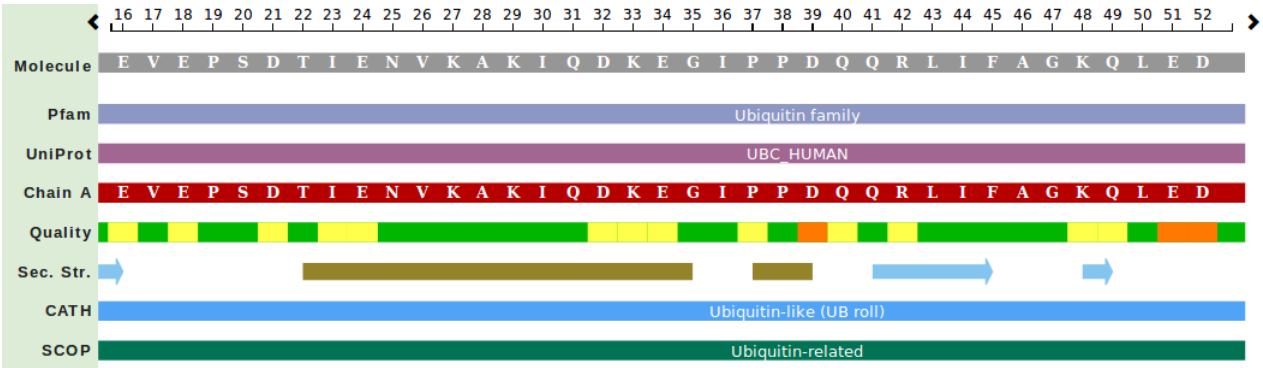


FIGURE 9 – Représentation annoté de 1ubi d'après PDBe (Protein Data Bank in Europe) avec le domaine défini par CATH et SCOP

Nous avons executé le programme original Protein Peeling 3D disponible en ligne [15].

(voir Figure 10).

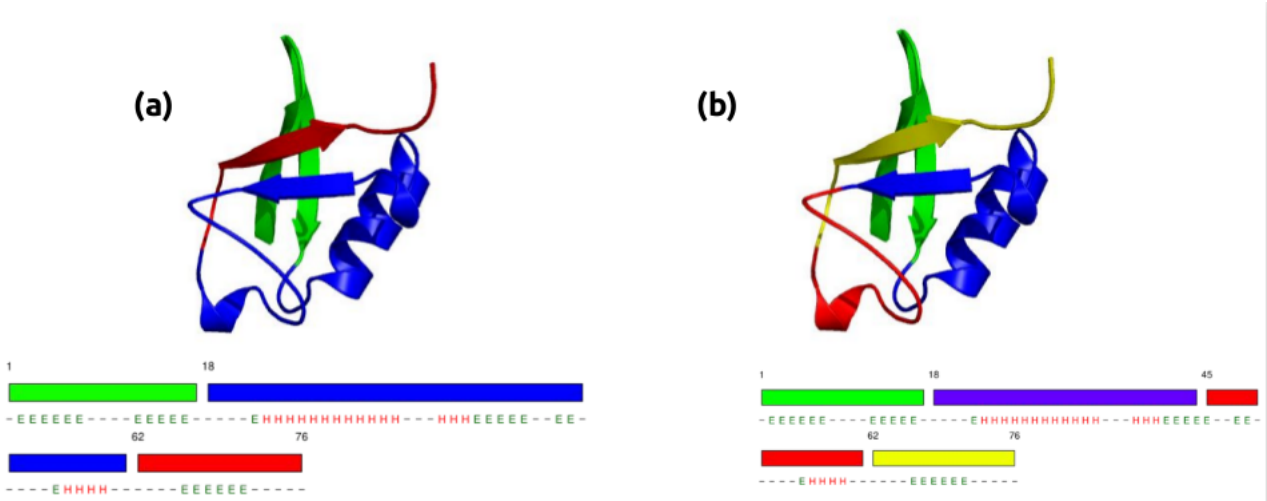


FIGURE 10 – Représentation de 1ubi avec le programme Protein Peeling d'origine
(a) Protéine découpée optimalement en 3 UPs avec sa séquence annoté selon les UPs et les structures secondaires.
(b) Protéine découpée optimalement en 4 UPs avec sa séquence annoté selon les UPs et les structures secondaires
Représentation structurale sur Pymol et des matrices sur Python

D'après la Figure 10 (a), le découpage en 3 UPs est défini par les intervalles sont : [1,17,61,76]. L'intervalle [18-61] étant moins favorable (critère d'énergie), elle a été divisé ; d'où le découpage en 4 UPs sur la Figure 10 (b).
Pour le découpage en 4 UPs les intervalles sont : [1,17,44,61,76]

Avec le programme implémenté pour ce projet nous retrouvons comme pour le Protein Peeling original le découpage en 3 UPs : [1, 21, 50, 77]. Cela est globalement similaire.
Quant au découpage en 4 UPs, il est retrouvé mais il s'avère peu favorable (pour le cas où min=10 et max=30). (voir Découpage n°2 et n°3 du Tableau 2). Cependant, la découpe [1, 10, 37, 61, 77] est cohérente comparée à celle des 4 UPs de Protein Pelling.
Pour les 5 UPs trouvés, nous n'avons pas assez d'information biologique pour discuter de la pertinence de ce découpage.

4.2 Comparaison de l'actine avec le résultat de Protein Peeling et la littérature

L'outil a décomposé l'actine (chaine A de 1atn) en 8 UPs.
En comparant avec les domaines défini par CATH (Figure 11), l'outil prédit 1 domaine de plus. CATH recense les domaines suivants :
- 1 : Nucléotidyltransferase - domain 5 [5-35]
- 2 : Actin- Chain 1 - domain 2 [36-70]

- 3 : Nucléotidyltransferase - domain 5 [72-135]
- 4 : Nucléotidyltransferase - domain 5 [137-182]
- 5 : Actin- Chain 1 - domain 4 [183-268]
- 6 : Nucléotidyltransferase - domain 5 [272-333]
- 7 : Nucléotidyltransferase - domain 5 [338-373]

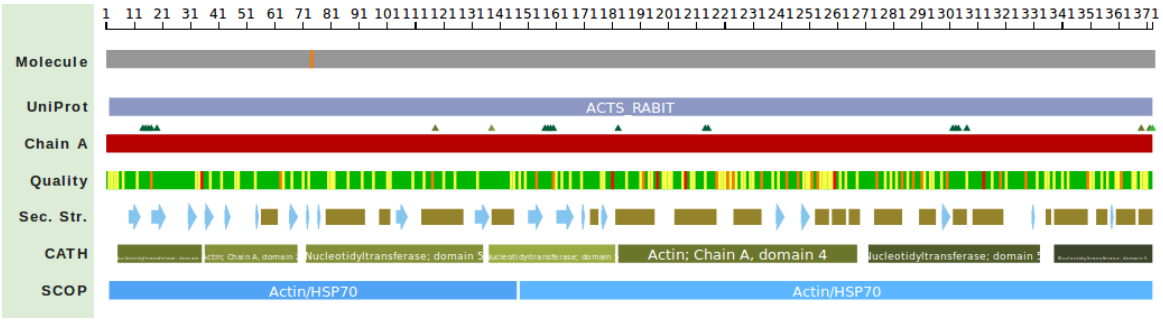


FIGURE 11 – Représentation annoté de 1atn d'après PDBe (Protein Data Bank in Europe) avec le domaine défini par CATH et SCOP

La décomposition trouvée par programme implémenté est : [1-50], [51-97], [98-146], [147-195], [196-244], [245-289], [289-334], [335-373].
Les deux derniers UPs de l'outil et de CATH sont similaires. Sinon, nous avons du mal a faire la comparaison pour le reste de la protéine.

Avec le programme Protein Peeling d'origine, le découpage au dernier degré montre également 8 domaines mais le découpage est différent, notamment au début de la protéine. En effet, du résidu 1 à 135, un seul PU est considéré contre 3 dans la version implémenté ici. Pour les résidus au milieu de la protéine, le découpage est différente. En revanche, les deux derniers UPs concordent avec le découpage proposé ainsi que celui de CATH. (voir Figure 11 et 12)

FIGURE 12 – Représentation de la séquence annoté de 1atn d'après le programme d'origine de Protein Peeling.

5 Conclusion

Nous avons donc parcouru dans les grandes lignes le fonctionnement du nouveau Protein Peeling implémenté pour le projet. Il réalise des découpages plausibles comme nous avons pu le voir sur deux cas : l'ubiquitine et l'actine.
Cependant, des améliorations sont encore nécessaires.
En effet, pour le découpage maximal, l'outil regarde l'index de partitionnement maximal (PI) pour un résidu donné. Cela restreint les combinaisons. Nous pouvons envisager de choisir plusieurs valeurs maximales de PI.
En outre, nous pourrions ajouter d'autres critères pour évaluer une UP.
Quant aux critères κ et σ, nous pourrions les normaliser en se basant sur le principe du Z-score (ie : couper la protéine un grand nombre de fois arbitrairement et calculer les moyennes κ etσ des découpages aléatoires).
Concernant la visualisation avec Pymol, le découpage des PDB n'est pas forcément la meilleure des solutions pour représenter des UPs, notamment les plus petites, car les structures secondaires peuvent ne pas être représenté sur Pymol.

Références

[1] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1) :235–242, January 2000.

[2] Tom L. Blundell, Bancinyane L. Sibanda, Rinaldo Wander Montalvão, Suzanne Brewerton, Vijayalakshmi Chelliah, Catherine L. Worth, Nicholas J. Harmer, Owen Davies, and David Burke. Structural biology and bioinformatics in drug design : opportunities and challenges for target identification and lead discovery. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 361(1467) :413–423, March 2006.

[3] Jane S. Richardson. The Anatomy and Taxonomy of Protein Structure. volume 34, pages 167–339. Academic Press, January 1981.

[4] B. L. Sibanda and J. M. Thornton. Beta-hairpin families in globular proteins. *Nature*, 316(6024) :170–174, July 1985.

[5] A. V. Efimov. Standard structures in proteins. *Progress in Biophysics and Molecular Biology*, 60(3) :201–239, January 1993.

[6] Nickolai Alexandrov and Ilya Shindyalov. Pdp : protein domain parser. *Bioinformatics*, 19(3) :429–430, 2003.

[7] Ganesan Pugalenthi, Govindaraju Archunan, and Ramanathan Sowdhamini. DIAL : a web-based server for the automatic identification of structural domains in proteins. *Nucleic Acids Research*, 33(Web Server issue) :W130–W132, July 2005.

[8] Jean-Christophe Gelly, Alexandre G. de Brevern, and Serge Hazout. 'Protein Peeling' : an approach for splitting a 3d protein structure into compact fragments. *Bioinformatics (Oxford, England)*, 22(2) :129–133, January 2006.

[9] W. Kabsch and C. Sander. Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12) :2577–2637, December 1983.

[10] Guillaume Postic, Yassine Ghouzam, Romain Chebrek, and Jean-Christophe Gelly. An ambiguity principle for assigning protein structural domains. *Science Advances*, 3(1), 2017.

[11] R Ramage, J Green, T W Muir, O M Ogunjobi, S Love, and K Shaw. Synthetic, structural and biological studies of the ubiquitin system : the total chemical synthesis of ubiquitin. *Biochemical Journal*, 299(Pt 1) :151–158, April 1994.

[12] Wolfgang Kabsch, Hans Georg Mannherz, Dietrich Suck, Emil F. Pai, and Kenneth C. Holmes. Atomic structure of the actin : DNase I complex. *Nature*, 347(6288) :37–44, September 1990.

[13] Natalie L. Dawson, Tony E. Lewis, Sayoni Das, Jonathan G. Lees, David Lee, Paul Ashford, Christine A. Orengo, and Ian Sillitoe. CATH : an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 45(Database issue) :D289–D295, January 2017.

[14] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP : a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4) :536–540, April 1995.

[15] Jean-Christophe Gelly and Alexandre G. de Brevern. Protein Peeling 3d : new tools for analyzing protein structures. *Bioinformatics (Oxford, England)*, 27(1) :132–133, January 2011.