**Rapport projet court : Assignation des structures secondaire de protéines**

**De Rita Eid M2BI**

**Introduction :**

La structure secondaire correspond aux repliements locaux de la chaine d’acide aminé d’une protéine. L’existence des structures secondaires vient du fait que les repliements énergétiquement favorables sont limités et que seule certaines conformations sont probables. C’est pourquoi une protéine peut être décrite par l’enchainement des acides aminés mais aussi de l’enchainement d’éléments de structures secondaires.

Parmi les conformations possibles :

Il y a conformation en hélice lorsque le squelette principal de la protéine adopte un repliement hélicoïdal périodique. Les liaisons hydrogènes, qui constitue cette configuration, se font entre des résidus consécutifs. Cette conformation est très fréquente dans les protéines. C‘est une structure compacte énergétique favorable.

Le brin beta est une structure périodique étendu, les liaisons hydrogènes qui le constitue se font entre des résidus distants. Un brin beta forme des liaisons d’hydrogène avec d’autre brin afin de se stabiliser, c’est les feuillets beta. Ces derniers ont une polarité qui vient du N terminal au C terminal. Nous avons donc les feuillets beta parallèles et les feuillets beta antiparallèles selon la polarite des 2 brins. En effet si cette dernière est la même on est face à un feuillet beta parallèle si non c’est un feuillet beta antiparallèle.

Dans cette étude, on a assez d’assigner des structures secondaires aux protéines. On veut donc savoir la structure secondaire des protéines en utilisant leur coordonnée. Les configurations recherchées seront les hélices, les feuillets beta parallèle et antiparallèles, et les turns. Pour cela on a implémenté la méthode de DSSP (*Define Secondary Structure of Proteins)* afin d’avoir un algorithme comparable à cette dernière.

**Méthode :**

Le programme DSSP a été conçu par Wolfgang Kabsch et Chris Sander pour normaliser l'attribution des structures secondaires. C’est une base de données d'affectations de structures secondaires étant donné les coordonnées de résolution atomique de la protéine soit pour toutes les entrées de protéines dans la Protein Data Bank (PDB). Ce programme se base sur l’étude des liaisons hydrogènes pour détecter des répétitions des motifs de liaison hydrogène élémentaires « tour » et « pont ». Ces derniers sont essentiels pour reconnaitre les structures secondaires. C’est pourquoi on a choisi cette méthode afin d’assigner les structures secondaires de nos protéines. En effet on implémente cet algorithme afin d’arriver à avoir un programme satisfaisant comparable à la méthode DSSP. Pour cela on utilise la même équation pour retrouver l’énergie favorable à la liaison hydrogène, c’est la suivante :

On calcule l'électrostatique énergie d'interaction entre deux groupes de liaison H en plaçant une partie chargée sur les atomes C,O (+q1, - q1) et N,H (-q2, + q 2 ), avec q1 et q2 étant et respectivement de valeurs -0,42e et 0,20e, e étant la charge électronique unitaire. En plus de F étant le facteur dimensionnel avec la valeur de 332kcal/mol. Et les termes {rAB} indiquent la distance entre les atomes A et B, prise parmi les atomes de carbone (C) et d'oxygène (O) du groupe C=O et l'azote (N) et des atomes d'hydrogène (H) du groupe NH. L’énergie obtenue est mesurée en kcal/mol. On considère qu’elle est favorable à une liaison hydrogène si cette dernière est inférieure à -0,5 kcal/mol.

Au début, on a déterminé les turns, on retrouve une liaison hydrogène au niveau :

En effet les turns sont formés grâce à des liaisons hydrogènes consécutives.

Puis pour assigner les hélices, on a choisi de définir qu’une succession de turns fait une hélice. On s’est focalisé sur les hélices alpha c’est pourquoi on n’a pas différencié les différents types d’hélice au niveau du code.

Alors que pour les brins beta, on a choisi de mesurer la distance entre les différents résidus qui forme une liaison hydrogène. Comme ces dernières doivent être entre 2 résidus distants et non consécutifs, on a déterminé qu’une liaison hydrogène qui apparait grâce à l’algorithme (et qui n’est pas entre 2 résidus consécutifs) est une liaison entre 2 brins beta.

Pour les feuillets parallèles, on regarde la différence entre les 2 résidus de la même liaison si elle est égale à la différence entre les 2 résidus de la liaison suivante. Alors que pour les feuillets antiparallèles la différence doit être significative c’est pourquoi on se base sur la différence entre les résidus de 2 liaisons hydrogènes consécutives s’ils sont proche ou non.

Le programme de cette étude est écrit en code python en utilisant plusieurs Library comme numpy, os, sys et PDBParser.

**Résultat :**

L’algorithme implémenté nous permet d’obtenir les structures secondaires des protéines. En effet on réussit à retrouver les différentes configurations qui se suivent de la même manière que dans la réalité. On arrive à retrouver les même paternes de structure secondaire.

Nous avons voulu comparer nos résultats à ceux du DSSP soit à l’algorithme de base. Pour cela on a lancé différentes protéines au niveau des 2 programmes et nous avons comparé les résultats des structures obtenues.

Choix des protéines pour la comparaison :

Nous avons choisi plusieurs protéines afin d’assigner leur structure secondaire. Le choix s’est principalement porté sur la longueur de la chaine polypeptidique mais aussi sur la variété des structures secondaires présente au niveau de cette protéine. En effet on commence avec la protéine 1bta qui possède des hélices et des feuillets beta parallèles puis on utilise la protéine 1mcy (la myoglobine) pour s’assurer qu’il définit bien les hélices. Puis on teste avec la protéine 1itv qui possède des feuillets antiparallèles pour assurer que notre algorithme détecte bien toutes les configurations voulues.

En premier temps, nous avons utilisé des petites protéines pour faire la comparaison. On remarque que pour ces protéines notre algorithme est très efficace pour assigner les structures secondaires avec un taux d’identification de 97% comparer à l’algorithme déjà existant de DSSP. On remarque que pour les hélices nous avons 100% d’identité entre les 2 algorithmes, donc notre programme détecte très bien les hélices. Les hélices sont facilement identifiables grâce à la périodicité de leur pattern. On observe une liaison hydrogène tous les 3,4 ou 5 résidus tout au long de cette structure secondaire. Le taux d’identification est le même soit 100% pour les feuillets parallèles qui sont efficacement identifiés. Ces feuillets sont directement définis avec la répétition de la même valeur dans la matrice de différence pour la liaison hydrogène. Les feuillets antiparallèles sont distingués aussi. Ces feuillets sont définis grâce à leurs liaisons hydrogènes qui se forment entre des résidus à sens opposé. Ces dernières étant plus difficiles à localisé uniquement grâce aux liaisons hydrogènes, nous avons un taux d’identification à 95%.

Pour les protéines de plus que 400 résidus, l’algorithme arrive à identifier facilement les hélices, elle reste avec un taux de 100% d’identification. Le feuillet parallèle reste distinguer avec un taux de 97%. En revanche pour les feuillets beta antiparallèles, ces derniers sont plus difficilement définis donc le taux d’indentification baisse jusqu’à 90%.

**Conclusion :**

L’algorithme utilisé dans cette étude donne des résultats presque identiques à celle du programme original DSSP. En effet, on retrouve des taux d’identité de très haute valeur ce qui prouve l’utilité de notre programme. La suite des structures secondaires au niveau des différentes protéine est définie et on retrouve les mêmes structures au niveau des résidus que ce soit avec la méthode DSSP ou avec notre algorithme. Il sert plus intéressant de développer l’algorithme encore plus afin qu’il puisse calculer les angles phi et psi.

**Reference**

W. Kabsch and C.Sander, Biopolymers 22 (1983) 2577-2637

Git :  <https://github.com/ritae9/DSSP_IMPL.git>