MARIAUX
Estelle

M2 BI

Université de Paris

Projet long

Accessibilité relative des résidus et évolution protéique

Responsable : Alexandre G. de Brevern

20 juillet 2021

Sommaire

Introduction……. 3

Matériels et méthodes…….. 4

Résultats et discussion……. 5

Conclusion……... 11

Introduction

Les protéines sont constituées d'une succession d'acides aminés, ces derniers constituent la structure primaire de la protéine. Ces acides aminés interagissent entre eux afin de former les structures secondaires comme les hélices alpha ou bien les feuillets bêta, ces structures interfèrent entre elles afin de former la structure tertiaire, aussi appelée structure tri-dimensionnelle qui donne sa forme à la protéine [1].

La structure tri-dimensionnelle de la protéine permet une certaine accessibilité relative des résidus, cette dernière peut aussi être assimilée à l'accessibilité aux solvants. Cette accessibilité est étroitement liée à la fonction de la protéine, en effet, en fonction de l'accessibilité de quelques résidus, une activité ou une fonction protéique pourront y être assimilées. Lorsque certains acides aminés, possédant une accessibilité plus élevée [2], sont en périphérie de la protéine, cela peut améliorer l'interaction de la protéine avec ses solvants. On pourra aussi noter que l'accessibilité des résidus peut être utilisée afin d'améliorer la prédiction des structures secondaires des protéines. En effet, les accessibilités des différents résidus constituant la protéine apportent des informations complémentaires afin de prédire plus précisément les structures de la protéine [3-4].

Il est souvent mal connu si, suite à l'évolution d'une protéine, l'accessibilité relative des résidus reste la même. L'accessibilité de deux acides aminés différents au sein d'une protéine n'est pas forcément la même et pourrait ainsi modifier l'accessibilité relative au cours de l'évolution de la protéine. La modification d'un ou plusieurs acides aminés pourraient donc aussi altérer les fonctions de la protéine durant son évolution.

Certaines études ont cherché à savoir s'il existe une relation entre l'accessibilité relative aux solvants d'une protéine et le taux d'évolution de celle-ci. L'article de Wilke et al. (2011) [5] évoque une relation linéaire mathématique entre ces deux valeurs, plus l'accessibilité de l'acide aminé est élevée, plus cet acide aminé sera susceptible de muter au cours de l'évolution de la protéine.

Dans ce projet, le but est dans un premier temps de calculer l'accessibilité relative des résidus pour une liste de fichiers PDB choisis. Pour chacun de ces fichiers PDB, plusieurs séquences avec un pourcentage d'identité élevé vont être sélectionnées afin de pouvoir calculer leur accessibilité relative. Ces séquences vont être considérées comme des modèles évolutifs, les différentes valeurs d'accessibilité aux solvants vont ainsi pouvoir être comparées avec la valeur d'accessibilité réelle, celle calculée pour le fichier PDB de base.

Matériels et méthodes

Afin de pouvoir étudier cette évolution de l'accessibilité, l'approche est constituée de plusieurs étapes décrites dans la figure 1.

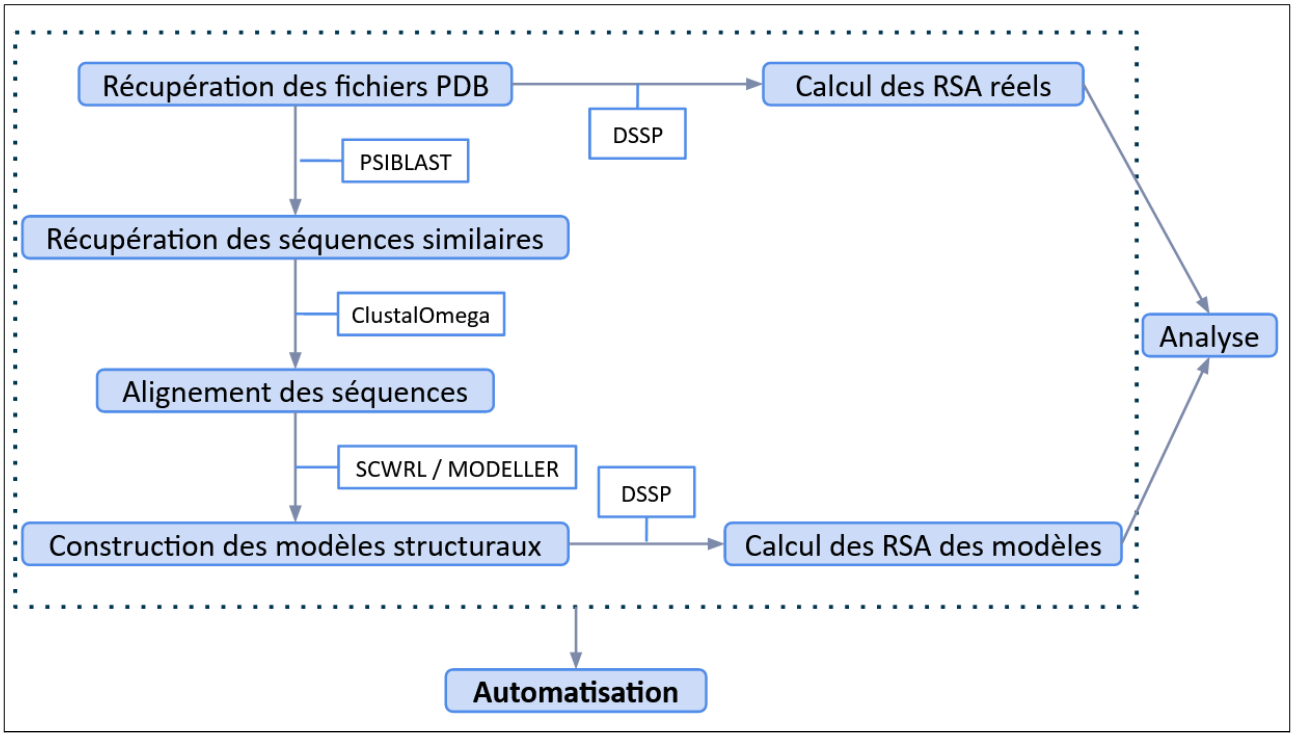

Figure 1. *Organisation des tâches du projet. RSA = relative solvent accessibility*

La première étape consiste ainsi à récupérer le fichier PDB de la protéine d'intérêt. Pour cela, le site RCSB (acronyme de l'anglais : *research collaboratory for structural bioinformatics*) propose un système de téléchargement de ses fichiers grâce à une ligne de commande permettant ainsi de récupérer les fichiers nécessaires pour la suite de l'étude, ainsi à partir d'une commande `wget`, le fichier PDB peut être téléchargé sur l'ordinateur.

Une fois le fichier récupéré, l'étape suivante est de calculer la surface accessible aux solvants. Afin de pouvoir réaliser cette étape, Biopython [6] a développé un module, issu du programme DSSP [7-8], permettant d'assigner automatiquement pour chaque acide aminé de la séquence protéique la structure secondaire prédite pour cet acide aminé, ce module `Bio.PDB.DSSP` [9] va en même temps calculer l'accessibilité relative de chaque acide aminé. Ce module peut utiliser la surface accessible (ASA = accessible surface area) de Miller et al. [10], de Sander et al. [11] ou de Tien et al. [12]. Dans notre étude, les valeurs utilisées sont celles de Miller et al., permettant de normaliser l'accessibilité aux solvants du résidu afin d'obtenir l'accessibilité relative aux solvants du résidu.

Ensuite, après avoir récupéré la séquence fasta de la protéine, une recherche de séquences similaires est réalisée à l'aide de l'outil BLAST [13] (*basic local alignment search tool*), disponible sur la plateforme du NCBI[1] (*national center for biotechnology information*). La réalisation d'une recherche à l'aide de l'outil PSI-BLAST [14] (*position-specific initiated BLAST*) va permettre de récupérer plusieurs séquences ayant un pourcentage d'identité élevé avec notre séquence d'intérêt. Les séquences qui sont intéressantes pourront être alignées entre elles à l'aide de ClustalOmega [15] afin de pouvoir comparer les différences d'acides aminés dans la suite de l'étude.

L'étape finale avant l'analyse statistique des données va être de récupérer les fichiers PDB des séquences homologues intéressantes pour cette étude et d'y appliquer le calcul d'accessibilité relative aux solvants. Ces résultats pourront être comparés aux résultats obtenus avec la séquence d'intérêt en réalisant différentes analyses, comme une moyenne de l'accessibilité des résidus entre les différentes séquences.

La langue de programmation choisie est en majorité du python car certains modules sont disponibles grâce à Biopython. Certains programmes vont être tournés en ligne bash. Voici le lien github du projet : ***https://github.com/estelleeeee/accessibility-prot-evol.git***

Résultats et discussion

Ce projet consiste en une automatisation de la recherche d'accessibilité des résidus et des modifications suite à une évolution protéique. Afin de travailler sur un grand nombre de protéine, le serveur PISCES permet l'accès à de nombreuses listes d'identifiants de protéines ainsi que leurs longueurs, les méthodes expérimentales utilisées (XRAY, NMR = *nuclear magnetic resonance*, etc.), leurs résolutions, etc. La liste choisie pour cette étude est la liste `cullpdb_pc20_res1.6_R0.25_d201103_chains3760`, elle permet d'accéder aux codes 4 lettres des fichiers PDB du RCSB ainsi que les chaînes sélectionnées par la liste. Un premier problème intervient à ce niveau là, certains fichiers ne sont pas forcément complets, le fichier 1A62.pdb par exemple, possède 5 acides aminés manquants de l'expérience, créant un fichier incomplet. Cela est problématique lors de la construction de la base de données des fichiers. De plus, certaines protéines possèdent plusieurs chaînes, cela ajoute de la difficulté pour le traitement des données, notamment pour le calcul de l'accessibilité. Si un fichier PDB possède plusieurs chaînes la fonction DSSP ne pourra plus fonctionner et l'accessibilité des résidus ne pourra plus être calculé. D'autres fichiers, trop volumineux, n'ont pas de fichiers PDB accessibles, seulement des fichiers XML ou bien mmCIF, causant cette fois-ci des problèmes d'automatisation de calcul.

L'étude est ici réalisée sur quelques fichiers, complets, sur une seule chaîne et étant assez petit pour être téléchargeable en format PDB. Pour la suite de l'étude, les fichiers PDB sont utilisés afin de calculer l'accessibilité relative. Deux exemples différents sont traités dans ce rapport, le fichier 1AHO.pdb et le fichier 1BTE.pdb (cf. tableau 1). Les fichiers sont tout d'abord récupérés

1 https://blast.ncbi.nlm.nih.gov/

par ligne de commande, à l'aide de système de téléchargement du RCSB. Les protéines sélectionnées par la liste PISCES n'ont pas forcément une unique chaîne, l'automatisation n'étant pas fonctionnelle, certains fichiers doivent être modifiés à la main afin de ne conserver qu'une seule chaîne peptidique pour permettre de réaliser le calcul de l'accessibilité aux solvants. Le module DSSP donne plusieurs informations sur les différents acides aminés, pour chacun d'eux, l'index d'identification est donné ainsi que l'identifiant de l'acide aminé, la structure secondaire associé à cet acide aminé, l'accessibilité relative, les angles phi et psi, et autres informations d'énergie entre les atomes (cf. tableau 2).

Code PDB	Taille (en AA)
1AHO	64
1BTE	97

Tableau 1. *Fichiers PDB choisis. AA = acides aminés*

Pour chacune des protéines, l'accessibilité de chaque acide aminé et la séquence fasta sont conservés en mémoire par le script python. A partir de ces séquences fasta, il est possible de lancer un PSI-BLAST. Pour chacune des séquences, l'algorithme récupère des séquences avec des pourcentages d'identité plus ou moins forts. Parmi ces séquences, celles qui ont un pourcentage d'identité de minimum 70 % sont retenues. Certaines possèdent des fichiers PDB, leurs structures sont résolues complètement, beaucoup d'autres n'ont pas de structures complètes et donc pas de fichier PDB associé. Par un soucis de temps, seules les séquences ayant un fichier PDB associé ont été retenues ici. Cependant, il est possible d'utiliser des logiciels comme MODELLER [17] afin d'obtenir des modèles structuraux pour ces séquences dont la structure n'est pas totalement résolue, et SCWRL4 [18] afin d'affiner les prédictions de conformation des chaînes latérales. Ici, ces étapes n'ont pas été réalisées.

```
(1, 'V', '-', 0.8309859154929577, 360.0, 162.7, 0, 0.0, 2, -0.3, 0, 0.0, 30, -0.0)
(2, 'K', 'E', 0.3024390243902439, -154.2, 160.8, 49, -1.8, 49, -2.8, 2, -0.0, 2, -0.3)
(3, 'D', 'E', 0.49079754601226994, -110.4, 154.4, -2, -0.3, 2, -0.3, 47, -0.2, 46, -0.2)
(4, 'G', 'E', 0.09523809523809523, 159.4, -176.4, 44, -2.4, 44, -2.7, -2, -0.3, 2, -0.7)
(5, 'Y', 'E', 0.2072072072072072, -81.8, 117.9, -2, -0.3, 53, -2.3, 42, -0.2, 42, -0.2)
(6, 'I', 'B', 0.0, -71.8, 147.7, 40, -0.8, 8, -0.8, -2, -0.7, 2, -0.3)
(7, 'V', 'B', 0.09154929577464789, -137.0, 161.8, 49, -1.3, 6, -0.2, 6, -0.2, 49, -0.2)
(8, 'D', '-', 0.4049079754601227, -76.9, -171.2, 4, -1.7, 3, -1.4, -2, -0.3, -1, -0.1)
(9, 'D', 'T', 0.7423312883435583, -72.9, -8.1, 1, -0.2, -1, -0.0, 2, -0.1, -2, -0.0)
(10, 'V', 'T', 0.676056338028169, -119.9, 24.4, 2, -0.1, -1, -0.2, 53, -0.0, -2, -0.0)
(11, 'N', 'S', 0.15286624203821655, 73.6, 23.0, -3, -1.4, 2, -0.5, 1, -0.2, 52, -0.1)
(12, 'C', '-', 0.15555555555555556, -99.3, 128.1, 50, -0.1, -4, -1.7, 46, -0.1, -1, -0.2)
(13, 'T', 'B', 0.18309859154929578, -68.5, 163.9, 50, -0.5, 2, -0.7, -2, -0.5, -6, -0.2)
(14, 'Y', '-', 0.26126126126126126, -91.2, 110.5, -8, -0.8, 32, -1.7, 32, -0.3, -1, -0.1)
(15, 'F', '-', 0.7715736040609137, -60.2, 145.9, -2, -0.7, 2, -0.3, 30, -0.2, 30, -0.2)
(16, 'C', '-', 0.022222222222222223, -155.8, 162.2, 28, -0.2, 3, -0.1, 21, -0.1, 28, -0.1)
(17, 'G', 'S', 0.36904761904761907, -125.8, -27.4, -2, -0.3, 2, -0.4, 1, -0.2, -1, -0.1)
(18, 'R', '-', 0.6532258064516129, -120.4, 144.7, 1, -0.1, 4, -1.4, 17, -0.1, 3, -0.4)
(19, 'N', 'H', 0.5859872611464968, -64.4, -38.0, -2, -0.4, 4, -2.6, 1, -0.2, 5, -0.2)
(20, 'A', 'H', 0.5943396226415094, -61.7, -36.8, 1, -0.2, 4, -2.0, 2, -0.2, -1, -0.2)
(21, 'Y', 'H', 0.3063063063063063, -59.9, -47.8, -3, -0.4, 4, -2.4, 2, -0.2, -1, -0.2)
(22, 'C', 'H', 0.0, -67.0, -36.7, -4, -1.4, 4, -2.8, 2, -0.2, 5, -0.2)
(23, 'N', 'H', 0.42038216560509556, -56.9, -44.0, -4, -2.6, 4, -2.0, 11, -0.3, -1, -0.2)
(24, 'E', 'H', 0.5515463917525774, -63.4, -48.1, -4, -2.0, 4, -2.1, 2, -0.2, -2, -0.2)
(25, 'E', 'H', 0.12886597938144329, -62.5, -39.2, -4, -2.4, 4, -1.0, 1, -0.2, -1, -0.2)
(26, 'C', 'H', 0.0, -65.7, -39.5, -4, -2.8, 5, -2.5, 1, -0.2, 3, -0.3)
(27, 'T', 'H', 0.44366197183098594, -74.5, -27.2, -4, -2.0, 3, -1.8, -5, -0.2, -2, -0.2)
```

```
(28, 'K', 'H', 0.6878048780487804, -59.2, -37.1, -4, -2.1, -1, -0.2, 1, -0.3, -2, -0.2)
(29, 'L', 'T', 0.3170731707317073, -86.5, 8.0, -4, -1.0, -1, -0.3, -3, -0.3, -2, -0.2)
(30, 'K', 'T', 0.44390243902439025, 74.9, 18.0, -3, -1.8, -3, -0.2, 1, -0.2, -2, -0.1)
(31, 'G', '-', 0.13095238095238096, -77.9, 170.9, -5, -2.5, -1, -0.2, -6, -0.2, 19, -0.2)
(32, 'E', 'S', 0.5721649484536082, -72.5, -49.0, 17, -2.3, 2, -0.3, 1, -0.3, 18, -0.2)
(33, 'S', 'E', 0.2153846153846154, -171.5, 174.4, 16, -1.0, 16, -2.7, -7, -0.1, 2, -0.3)
(34, 'G', 'E', 0.11904761904761904, -173.3, -179.7, -2, -0.3, -11, -0.3, 14, -0.3, 2, -0.3)
(35, 'Y', 'E', 0.27927927927927926, -154.3, 169.4, 12, -2.2, 12, -2.6, -2, -0.3, 2, -0.7)
(36, 'C', 'E', 0.037037037037037035, -92.4, 110.8, -2, -0.3, 2, -1.0, 10, -0.2, 10, -0.2)
(37, 'Q', 'E', 0.29797979797979796, -82.8, 106.0, 8, -2.7, 8, -2.1, -2, -0.7, 3, -0.8)
(38, 'W', 'E', 0.6696035242290749, -73.9, 143.3, -2, -1.0, 6, -0.1, 1, -0.2, 5, -0.1)
(39, 'A', 'E', 0.8113207547169812, 57.0, 41.2, -2, -0.2, -1, -0.2, 1, -0.1, 5, -0.2)
(40, 'S', 'E', 0.09230769230769231, -71.9, 173.0, 3, -1.2, 3, -2.2, -3, -0.8, -1, -0.1)
(41, 'P', 'T', 0.8308823529411765, -61.7, -12.7, 0, 0.0, 3, -0.1, 0, 0.0, -1, -0.1)
(42, 'Y', 'T', 0.38738738738738737, -109.7, 14.6, 1, -0.6, 2, -0.2, 20, -0.0, -29, -0.1)
(43, 'G', 'E', 0.36904761904761907, 85.0, -165.4, -3, -2.2, -3, -1.2, 2, -0.1, -1, -0.6)
(44, 'N', 'E', 0.17834394904458598, -65.2, 131.0, -5, -0.2, 2, -0.3, -2, -0.2, -6, -0.2)
(45, 'A', 'E', 0.0, -150.9, 143.3, -8, -2.1, -8, -2.7, -2, -0.3, 2, -0.2)
(46, 'C', 'E', 0.0, -72.7, 136.6, -32, -1.7, -40, -0.8, -2, -0.3, 2, -0.4)
(47, 'Y', 'E', 0.23873873873873874, -120.0, 131.7, -12, -2.6, -12, -2.2, -2, -0.2, 2, -0.3)
(48, 'C', 'E', 0.0, -119.7, 150.1, -44, -2.7, -44, -2.4, -2, -0.4, 2, -0.4)
(49, 'Y', 'E', 0.3918918918918919, -106.7, 140.5, -16, -2.7, -17, -2.3, -2, -0.3, -16, -1.0)
(50, 'K', 'E', 0.4292682926829268, 57.1, 47.6, -2, -0.4, -47, -0.2, -19, -0.2, -1, -0.2)
(51, 'L', 'E', 0.0, -72.6, 151.8, -49, -2.8, -49, -1.8, -3, -0.3, -1, -0.2)
(52, 'P', '-', 0.18382352941176472, -60.0, 153.8, 0, 0.0, 3, -1.8, 0, 0.0, -1, -0.1)
(53, 'D', 'T', 0.7239263803680982, -64.9, -21.2, 1, -0.3, -2, -0.1, -3, -0.1, -3, -0.0)
(54, 'H', 'T', 0.875, -79.9, -12.3, 2, -0.1, -1, -0.3, 0, 0.0, 2, -0.2)
(55, 'V', 'S', 0.1267605633802817, -77.1, 136.0, -3, -1.8, 2, -0.4, -26, -0.1, -4, -0.1)
(56, 'R', '-', 0.6290322580645161, -76.7, 132.0, -2, -0.2, -49, -1.3, -49, -0.2, 2, -0.3)
(57, 'T', 'B', 0.21830985915492956, -117.3, 168.5, -2, -0.4, -51, -0.2, -51, -0.2, -53, -0.0)
(58, 'K', '-', 0.16097560975609757, -58.2, 133.5, -53, -2.3, -53, -0.1, -2, -0.3, -1, -0.1)
(59, 'G', '-', 0.3333333333333333, -123.5, -175.9, 1, -0.2, -1, -0.1, 2, -0.2, -3, -0.0)
(60, 'P', 'S', 1.0, -63.7, 149.0, 0, 0.0, -1, -0.2, 0, 0.0, 2, -0.1)
(61, 'G', 'S', 0.6190476190476191, 115.7, 135.8, -3, -0.1, 2, -0.3, 2, -0.0, -2, -0.2)
(62, 'R', '-', 0.8548387096774194, -87.0, 138.3, -2, -0.1, 2, -0.3, -50, -0.0, -50, -0.1)
(63, 'C', '-', 0.2518518518518518, -150.5, 118.4, -2, -0.3, -50, -0.5, -52, -0.1, -2, -0.0)
(64, 'H', '-', 1.0, -131.2, 360.0, -2, -0.3, -1, -0.0, -52, -0.1, -22, -0.0)
```

Tableau 2. *Résultats de DSSP de la protéine 1AHO.pdb. Légende des structures secondaires : '-' :* none, *'E' :* strand, *'B' :* isolated beta-bridge residue, *'T' :* turn, *'S' :* bend, *'H' :* alpha helix (4-12)

Il a donc été possible de sélectionner 3 protéines ayant un pourcentage d'identité fort avec la protéine 1AHO (cf. tableau 3.A). Ces 3 protéines sont récupérées afin de pouvoir poursuivre les analyses de l'accessibilité relative des résidus au cours de l'évolution protéique. Ces protéines ont aussi été alignées avec ClustalW (cf. tableau 3.B).

Code PDB	**Pourcentage d'identité**
4AEI	100 %
1SEG	86 %
2KBH	73 %

__A.__ Informations des fichiers sélectionnés

```
CLUSTAL O(1.2.4) multiple sequence alignment


2KBH_A      VKDGYIADDRNCPYFCGRNAYCDGECKKNRAESGYCQWASKYGNACWCYKLPDDARIMKP    60
1AHO_A      VKDGYIVDDVNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYKLPDHVRTKGP    60
4AEI_A      VKDGYIVDDVNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYKLPDHVRTKGP    60
1SEG_A      VKDGYIVKNYNCTYFCFRNAYCNEECTKLKGESGYCQWASPYGNACYCYKLPDHVPIRVP    60
            ******..: ** *** *****: **.* :.********* *****:******..    *


2KBH_A      GRCNGG  66
1AHO_A      GRCH--  64
4AEI_A      GRCHX-  65
1SEG_A      GKCH--  64
            *:*:
```

__B.__ Alignement des séquences par ClustalW

Tableau 3. *Séquences homologues au fichier 1AHO.pdb*

Il est possible de remarquer dans cet alignement (tab. 3.B) que dans cette situation, les protéines n'ont pas d'insertions ou de délétions au milieu des séquences. Cela permet de faciliter l'analyse, on pourra cependant remarquer que la protéine 2KBH possède 2 glycines supplémentaires en C-terminal.

Afin d'analyser encore un peu plus les différentes séquences, il est possible de réaliser des comparaisons des accessibilités entre les séquences. Trois séquences sont de la même taille, il a donc été possible de réaliser une moyenne des accessibilités des séquences 1AHO, 4AEI et 1SEG. Un graphique (cf. figure 2) montre cette moyenne (*average*) ainsi que les accessibilités des différents acides aminés, on pourra aussi noter que dans une autre version du graphique (cf. figure 3), l'accessibilité de la protéine 2KBH est aussi affichée mais n'est pas prise en compte pour le calcul de la moyenne d'accessibilité.

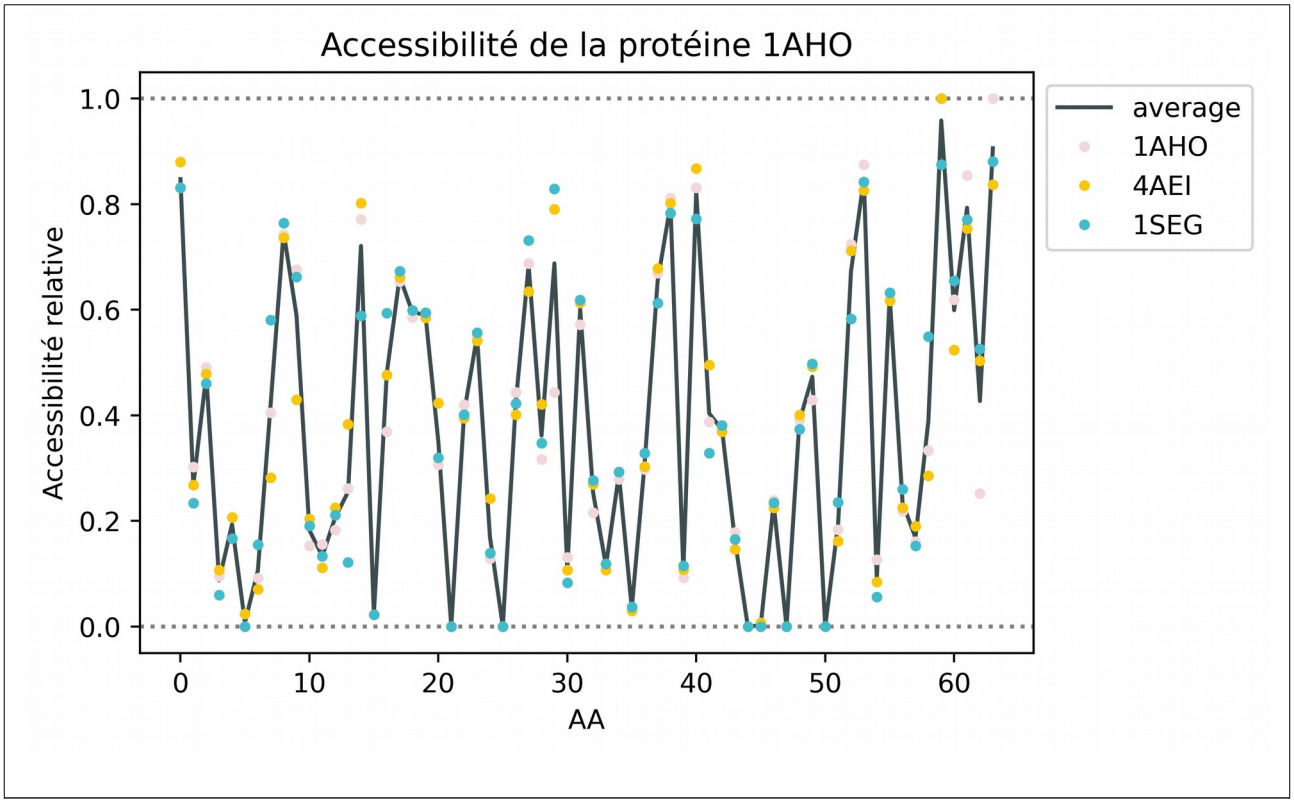

Figure 2. *Accessibilité moyenne des protéines 1AHO, 4AEI et 1SEG. AA = Acides Aminés*

Ces graphiques permettent ainsi d'observer les différences entre les séquences visuellement entre les différentes positions. Pour la majorité des acides aminés, malgré quelques différences, les accessibilités restent globalement les mêmes, en notant que même à acide aminé égal et position égal les accessibilités diffèrent légèrement. Pour certaines positions les accessibilités restent exactement les mêmes, lorsque l'acide aminé n'est pas du tout accessible (accessibilité égale à 0) pour la position 22 (cytosine), la position 26 (cytosine, uniquement en fig. 2), la position 45 (alanine), la position 48 (cytosine) et enfin la position 51 (leucine). Les autres positions diffèrent souvent de très peu, mais certaines positions ont de plus grandes différences comme la 8 avec un

changement de l'aspartate en lysine ou bien la position 30 pourtant une position sans changement d'acide aminé.

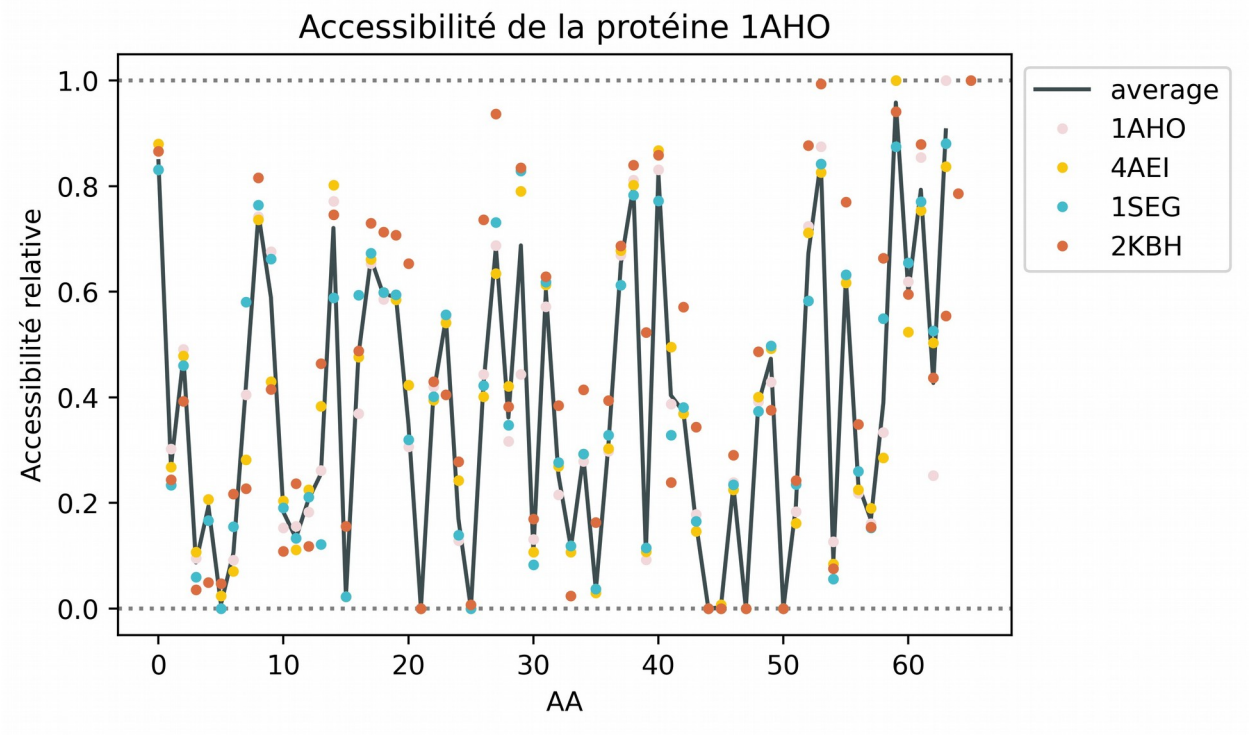

Figure 3. *Accessibilité moyenne des protéines 1AHO, 4AEI et 1SEG et accessibilité de 2KBH. AA = Acides Aminés.*

Il est possible de réaliser le même type d'analyse pour la séquence 1BTE, de la même manière que précédemment, 2 séquences ont pu être sélectionnées avec un fort pourcentage d'identité avec notre séquence cible, la chaîne B de la protéine 1LX5 et la chaîne A de la protéine 5NH3 (cf. tableau 4.A). Ces séquences ont pu être alignées avec ClustalW (cf. tableau 4.B).

Code PDB	**Pourcentage d'identité**
1LX5	95 %
5NH3	94 %

A. *Informations des fichiers sélectionnés*

B. *Alignement des séquences par ClustalW*

```
CLUSTAL O(1.2.4) multiple sequence alignment


1BTE_A      -----SETQECLFFNANWERDRTNQTGVEPCYG----RRHCFATWKNISGSIEIVKQGCW    51
1LX5_B      AILGRSETQECLFFNANWERDRTNQTGVEPCYGDKDKRRHCFATWKNISGSIEIVKQGCW    60
5NH3_A      AILGRSETQECLFFNANWEKDRTNQTGVEPCYGDKDKRRHCFATWKNISGSIEIVKQGCW    60
                 **************:*************    **********************


1BTE_A      LDDINCYDRTDCIEKKDSPEVYFCCCEGNMCNEKFSYFPEME------------------    93
1LX5_B      LDDINCYDRTDCIEKKDSPEVYFCCCEGNMCNEKFSYFPEME------------------    102
5NH3_A      LDDINCYDRTDCVEKKDSPEVYFCCCEGNMCNEKFSYFPEMEVTQPTSNPVTPKPPEFRH    120
            ************:******************************


1BTE_A      --  93
1LX5_B      --  102
5NH3_A      DS  122
```

Tableau 4. *Séquences homologues au fichier 1BTE.pdb*

La première remarque à faire par rapport à l'étude avec 1BTE est que cette fois-ci, la séquence possède une insertion lors de l'alignement. Cette insertion va compliquer l'étude de notre alignement, en effet une moyenne ne pourra pas être correctement calculée sur cet alignement. Il est cependant possible de modifier les données qui permettent d'obtenir les graphiques, ainsi 4 valeurs nulles ont été ajoutées aux données correspondant au gap dans la séquence. Il pourra être noté qu'une valeur nulle a aussi été ajoutée au début des séquences 1BTE et 1LX5 car la séquence 5NH3 avait une serine en début de séquence qui n'était pas présente dans les autres.

Il est aussi intéressant de faire remarquer que les PDB récupérés ne correspondent pas forcément à l'alignement obtenu avec ClustalW suite aux séquences obtenues par PSI-BLAST. En effet, sur ClustalW, la séquence 1BTE fait 93 AA (acides aminés), 1LX5 fait 102 AA tandis que 5NH3 fait 122 AA alors qu'avec les fichiers PDB récupérés nous obtenons respectivement, 91, 94 et 94 AA.

Après avoir adapté les données pour que les séquences soient bien alignées sur le graphique, il est possible d'afficher les données (cf. figure 4). Dans l'étude de 1BTE, il n'a pas été possible de calculer la moyenne dû à l'insertion dans les séquences 1LX5 et 5NH3 par rapport à la séquence 1BTE.

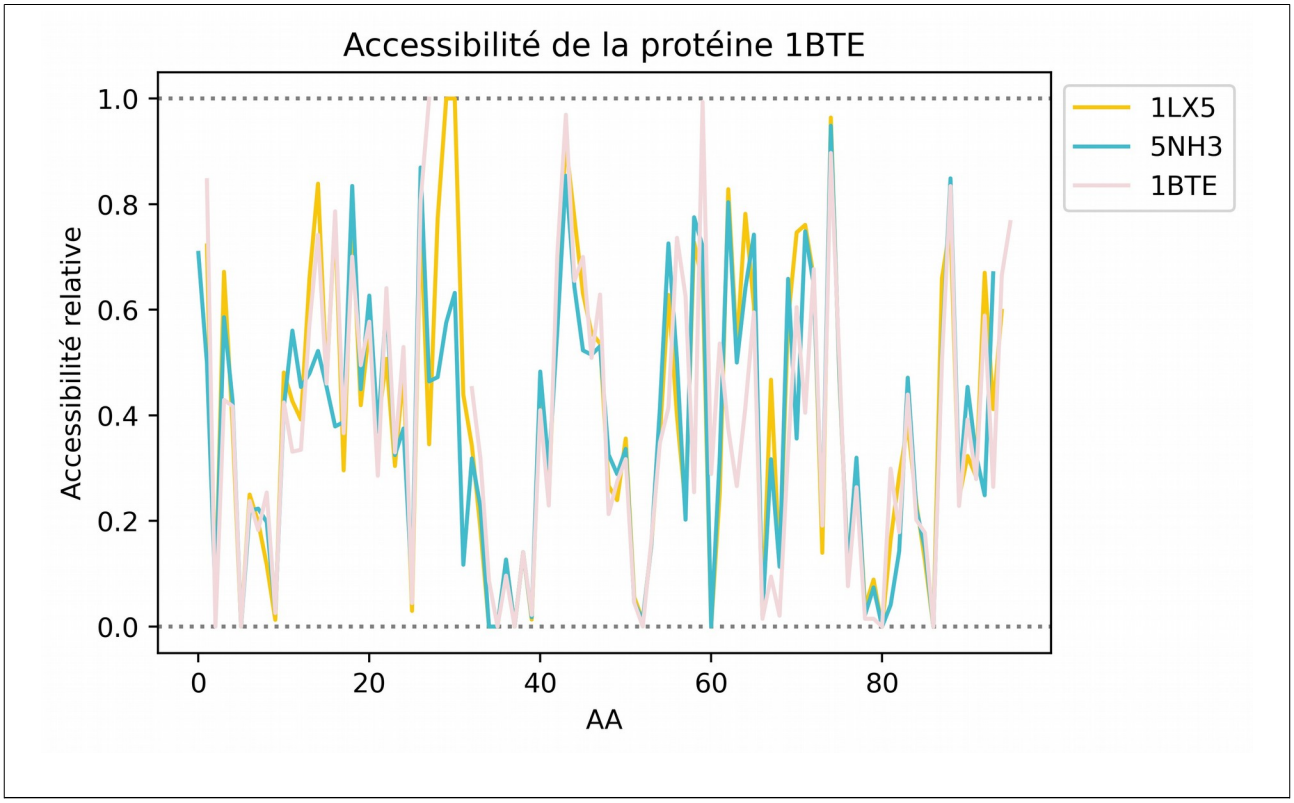

***Figure 4**. Accessibilité des protéines 1BTE, 1LX5 et 1BTE. AA = Acide Aminé*

La figure ci-dessus, nous montre que l'accessibilité se conserve assez bien dans l'ensemble. Il pourra être noté que malgré la correspondance des acides aminés dans les insertions pour les

séquences 1LX5 et 5NH3, l'accessibilité de cette zone semble très différente entre les deux séquences. Des accessibilités équivalentes peuvent être retrouvées dans cette étude aussi, toujours pour des accessibilités des acides aminés à 0.

Conclusion

Malgré les deux semaines complémentaires afin d'améliorer ce projet, les analyses restent encore une fois trop maigres afin de tirer des conclusions sur ce projet. La principale erreur réalisée lors de la première ébauche du projet était d'avoir traiter l'accessibilité relative des résidus comme étant une accessibilité globale alors que cette dernière doit être traitée localement. Ce problème a été résolu avec cette nouvelle version. Il a été possible de comparer l'accessibilité relative du fichier PDB de base avec les fichiers modèles et d'observer les différences entre les acides aminés.

Cependant seulement 2 fichiers PDB ont été analysés ici et une grande partie du projet consistait à l'automatisation des recherches. De nombreux problèmes peuvent se corriger à la main lors de petites analyses (comme le réglage des insertions et délétions au sein des alignements des séquences) mais ne peuvent pas être permises lors du traitement de 3760 chaînes protéiques.

Des analyses complémentaires auraient pu être réalisées avec ce projet, mais le projet n'est toujours pas arrivé à bout. L'aspect automatisation et manque de moyen pour traiter les données localement sur l'ordinateur (difficulté à faire marcher PSI-BLAST, ClustalW localement) ont ajouter de la difficulté à la réalisation de ce projet.

Bibliographie

1. C.B. Anfinsen ; Principles that govern the folding of protein chains. *Sciences* (1973).

2. L. Lins, A. Thomas, R. Brasseur ; Analysis of accessible surface of residues in proteins. *Protein science* (2003).

3. R. Adamczak, A. Porollo, J. Meller ; Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins. *Proteins : Structure, Function, and Bioinformatics* (2005).

4. A. Momen-Roknabadi, M. Sadeghi, H. Pezeshk, S.A. Marashi ; Impact of residue accessible surface area on the prediction of protein secondary structures. *BMC Bioinformatics* (2008).

5. D.C. Ramsey, M.P. Scherrer, T. Zhou, C.O. Wilke ; The relationship between relative accessibility and evolutionary rate in protein evolution. *Genetics* (2011).

6. P.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J.L. de Hoon ; Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009)

7. W.G. Touw, C. Baakman, J. Black, T.A.H. Te Beek, E. Krieger, R.P. Joosten, G. Vriend ; A series of PDB related databases for everyday needs. *Nucleic Acids Research* (2015).

8. W. Kabsch, C. Sander ; Dictionary of protein secondary structure: pattern recognition of hydrogen-boded and geometrical features. *Biopolymers* (1983).

9. T. Hamelryck, B. Manderick ; PDB file parser and structure class implemented in Python. *Bioinformatics* (2003).

10. S. Miller, J. Janin, A.M. Lesk, C. Chothia ; Interior and Surface of Monomeric Proteins. *Journal of Molecular Biology* (1987).

11. B. Rost, C. Sander ; Conservation and prediction of solvent accessibility in protein families. *Proteins : Structure, Function, and Genetics* (1994).

12. M.Z. Tien, A.G. Meyer, D.K. Sydykova, S.J. Spielman, C.O. Wilke ; Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS ONE* (2013).

13. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman ; Basic local alignment search tool. *Journal of Molecular Biology* (1990).

14. S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman : Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Ressearch* (1997).

15. F. Madeira, Y.M. Park, J. Lee, et al ; The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research* (2019).

16. G. Wang, R.L. Dunbrack, Jr. ; PISCES: a protein sequence culling server. *Bioinformatics* (2003).

17. A. Šali, T.L. Blundell ; Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* (1993).

18. G.G. Krivov, M.V. Shapovalov, R.L. Dunbrack, Jr. ; Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* (2009)