MARIAUX
Estelle

M2 BI

Université de Paris

Projet long

Accessibilité relative des résidus et évolution protéique

Responsable : Alexandre G. de Brevern

19 mars 2021

Sommaire

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.

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) [3] 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 comparer 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. La première étape consiste à récupérer le fichier PDB de la protéine d'intérêt. Pour cela, le site RCSB[1] (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.

Une fois le fichier récupéré, l'étape suivante est de calculer la surface accessible aux solvants. Afin de pouvoir réaliser cette étape, Biopython [4] a développé un module, issu du programme DSSP [5-6], 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* [7] va en même temps calculer l'accessibilité relative de chaque acide aminé.

Ensuite, après avoir récupéré la séquence fasta de la protéine, une recherche de séquences similaires est réalisée à l'aide de l'outil BLAST [8] (*basic local alignment search tool*), disponible sur la plateforme du NCBI[2] (*national center for biotechnology information*). La réalisation d'une recherche à l'aide de l'outil PSI-BLAST [9] (*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 [10] 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.

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.

Malheureusement, ce projet n'est pas fini. Voici cependant, le lien du github du projet : ***https://github.com/estelleeeee/accessibility-prot-evol.git***

1 https://www.rcsb.org/
2 https://blast.ncbi.nlm.nih.gov/

Résultats et discussion

Comme précisé quelques lignes ci-dessus, ce projet n'est pas terminé. J'en suis désolée. Cependant voici les résultats obtenus, les problèmes rencontrés, les solutions et les futures pistes à tester.

Ce projet devait aboutir à une automatisation de cette recherche d'accessibilité des résidus et ses possibles modifications suite à l'évolution de la protéine. Cette automatisation nécessitait une longue liste d'identifiants PDB de protéines, elle peut être récupérée sur le serveur PISCES [11]. Ce serveur permet un accès à de nombreuses listes d'identifiants de protéines ainsi qu'à leurs longueurs, les méthodes expérimentales utilisées (XRAY, NMR = *nuclear magnetic resonance*, etc.), leurs résolutions, etc. Cette liste permet d'accéder aux codes 4 lettres des fichiers PDB du RCSB. Le premier problème apparaît dès cette étape de récupération des fichiers PDB. Certains ne sont pas complets, par exemple le fichier 1A62.PDB comporte 5 acides aminés manquants de l'expérience, créant un fichier incomplet. Cela pose des problèmes dans la construction de la base de données des fichiers. De plus, les fichiers possédant plusieurs chaînes sont difficiles à gérer pour la suite de l'étude, notamment pour le calcul de l'accessibilité. 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.

Cependant, après sélection de quelques fichiers, complets, ayant une seule chaîne et étant assez petit pour être téléchargeable en format PDB, la suite de l'étude a pu être réalisée. Pour ces fichiers, l'accessibilité relative est calculée. Deux exemples vont être traités dans ce rapport, le fichier 1AHO.pdb et le fichier 1BTE.pdb (cf. tableau 1).

Code PDB	Taille (en AA)	RSA
1AHO	64	23,91
1BTE	97	26,27

Tableau 1. *Fichiers PDB choisis. AA = acides aminés ; RSA = relative solvent accessibility : accessibilité relative au solvant*

Les fichiers sont récupérés par ligne de commande, il existe quelques problèmes concernant la récupération de ces fichiers. L'automatisation n'est pas fonctionnelle pour la suite de cette étude, certains fichiers ont dû être altérés afin de ne conserver qu'une seule chaîne polypeptidique pour le calcul de l'accessibilité au solvant. Cette accessibilité, indiquée dans la troisième colonne du tableau 1 est calculée à partir des données obtenues grâce au module DSSP. Le module va donner l'accessibilité relative aux solvants pour chacun des acides aminés de la protéine, la valeur affichée dans cette troisième colonne est l'addition de toutes ces valeurs.

Après avoir récupéré les séquences fasta des protéines d'intérêt, un PSI-BLAST peut être lancé. Pour chacune des séquences, l'algorithme récupère des séquences avec des pourcentages d'identité plus ou moins fort. Parmi ces séquences, celles qui ont un pourcentage d'identité de minimum 70 % sont retenues. Certaines de ces séquences 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 [12] afin d'obtenir des modèles structuraux pour ces séquences dont la structure n'est pas totalement résolue, et SCWRL4 [13] afin d'affiner les prédictions de conformation des chaînes latérales. Ici, ces étapes n'ont pas été réalisées.

Pour chaque fichiers récupérés pour l'exemple de cet étude (1AHO.pdb et 1BTE.pdb), plusieurs fichiers PDB (tableau 2.A. et tableau 3.A. respectivement) vont être récupérés afin de pouvoir poursuivre l'analyse de l'accessibilité relative des résidus au cours de l'évolution protéique. Afin d'avoir une meilleure vision des différentes séquences, elles sont alignées entre elles dans les tableaux 2.B. et 3.B. pour les analyses de 1AHO.pdb et 1BTE.pdb respectivement.

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 2. *Séquences homologues au fichier 1AHO.pdb*

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

A. *Informations des fichiers sélectionnés*

```
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
```

B. *Alignement des séquences par ClustalW*

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

Pour chaque fichiers récupérés lors de cette étape, le calcul d'accessibilité relative aux solvants va être appliqué afin d'observer les éventuelles différences. Les résultats de ces calculs sont affichés dans les tableaux 4 et 5.

Code PDB	Taille (en AA)	RSA
1AHO	64	23,91
4AEI	65	24,44
1SEG	64	24,43
2KBH	66	29,15

Tableau 4. *Données pour 1AHO.pdb*

En comparant les résultats obtenus dans le tableau 4, notamment la colonne de l'accessibilité aux solvants, avec les alignements du tableau 2.B., certains éléments peuvent être remarqués. Tout d'abord le fait que les protéines 4AEI et 1SEG ont une accessibilité relative aux solvants presque identiques, cependant, l'une a un pourcentage d'identité de 100 % mais un acide aminé en plus dans sa séquence tandis que l'autre a un pourcentage d'identité de 86 % mais une séquence de la même taille que la séquence cible. Les résultats de la protéine 2KBH semble rester cohérent, la protéine est très légèrement plus longue et possède plus de modifications dans sa séquence, un changement plus important dans l'accessibilité n'est donc pas forcément étonnant. Ces résultats restent malgré tout à étudier plus en profondeur car le module DSSP ne compte pas forcément tous les acides aminés du fichier PDB, causant des problèmes de calcul. En effet, pour le fichier 4AEI.pdb, la séquence affiche un acide aminé de plus que la séquence d'intérêt, cependant lors du calcul de l'accessibilité relative, cet acide aminé n'apparaît pas. Il est cependant possible de penser que la nature de l'acide aminé modifié dans les séquences pourra avoir un impact sur la conformation de la protéine (avec un acide aminé plus volumineux ou bien au contraire un plus petit) et ce changement de conformation modifiera l'accessibilité des résidus de la protéine.

Il est possible de faire le même raisonnement avec la protéine 1BTE, plusieurs protéines ont été identifiées comme ayant un pourcentage élevé d'identités entre les séquences. Ces protéines sont étudiées dans le tableau 5.

Code PDB	Taille (en AA)	RSA
1BTE	97	34,53
1LX5	102	37,77
5NH3	122	35,50

Tableau 5. *Données pour 1BTE.pdb*

Ce tableau nous apporte des informations complémentaires, ici les séquences ont des longueurs avec plus de différences que dans le cas précédent. Cependant, l'accessibilité relative ne semble pas être modifié par cette grande différence de taille, les protéines 1BTE et 5NH3, pourtant

différentes de 25 acides aminés, ont des accessibilités relatives assez proches contrairement à la protéine 1LX5. Ces différences pourraient être l'origine de modifications particulières sur l'une des séquences, provoquant de grands changements de conformation dans la protéine. Pourtant les chaînes d'intérêt des protéines 1LX5 et 5NH3 ont pratiquement les mêmes modifications (à défaut de la prolongation sur 19 acides aminés de 5NH3 à l'extrémité C-terminale), ces différences soulèvent un nouveau problème non réglé. Lors du calcul de cette accessibilité, malgré la présence des différents atomes composant les acides aminés de l'extrémité C-terminale de 5NH3 dans le fichier PDB, ces acides aminés ne sont pas inclus dans le calcul de l'accessibilité relative.

Conclusion

Les analyses restent cependant trop maigres pour tirer de vraies conclusions sur ce projet. La partie d'analyses statistiques ne possèdent pas assez de résultats, l'idée avait été de comparer l'accessibilité relative du fichier PDB de base avec les fichiers modèles et d'observer une corrélation possible. Cependant seulement 2 fichier PDB ont été analysés ici et certains problèmes sur ces études ne sont pas réglés, cela ne permet pas d'aller jusqu'au bout de ces analyses.

De nombreuses choses auraient été à faire avec ce projet, cependant par manque de temps et de motivation, je n'ai pas réussi à mener le projet à bout, en dépit de l'intérêt apporté au sujet. En reprenant le projet depuis le début, sous une différente approche afin de ne pas être bloquée par les différents formats de fichiers, par les versions locales de certains logiciels, par la gestion du temps, etc., cela aurait pu avoir des résultats intéressants.

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

4. 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)

5. 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).

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

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

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

9. 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).

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

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

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

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