https://github.com/meryambyt/Projet_Threading.git
Meryam BOULAYAT
M2 Bioinformatique
71704209

Université Paris Cité

Projet Court

Élaboration d'un outil permettant l'alignement entre une séquence et une structure protéique par double programmation dynamique

---

Programmation 3 et projet tuteuré

Année 2022/2023

Introduction

Les protéines jouent un rôle fondamental dans une multitude de processus biologiques. Ainsi, la compréhension de leur structure tridimensionnelle est essentielle pour comprendre les mécanismes sous-jacents de leur fonction. Cependant, déterminer expérimentalement la structure de chaque protéine est une tâche consommatrice de temps et de ressources. En conséquence. De plus, malgré les avancées technologiques, la plupart des protéines restent non caractérisées sur le plan structural. Ainsi, il existe des méthodes telles que l'alignement structure-séquence (enfilage ou threading), permettant de prédire la structure tridimensionnelle d'une protéine à partir de sa séquence d'acides aminés. Plutôt que de s'appuyer sur des données expérimentales, l'enfilage utilise des modèles statistiques et des informations issues de bases de données de structures connues pour estimer la conformation spatiale d'une protéine donnée. Dans le cadre de ce projet, nous nous sommes basés sur l'outil THREADER de David Jones afin de réaliser un programme capable de réaliser un alignement optimal entre une structure et une séquence en utilisant de la double programmation dynamique.

Matériels et méthodes

La conception de l'algorithme d'enfilage a été basée sur l'outil THREADER de David Jones. Ainsi, à partir d'une séquence protéique au format FASTA et d'une structure PDB (Protein Data Bank), des matrices d'alignements ont été réalisées afin d'identifier l'alignement optimal. Pour cela, la méthode de double programmation dynamique a été utilisée. Celle-ci permet la construction de matrices de bas niveau (nommées L) et une de haut niveau (nommée H). Une matrice L est construite en fixant une position donnée (i,j) où i représente un acide aminé de la séquence et j une position dans la structure. Le score optimal de la matrice de bas niveau est ensuite calculé par le biais de l'algorithme de Needleman and Wunsch (NW) et est reporté dans la case H(i,j). Une fois la matrice H construite, celle-ci est utilisée comme matrice de score pour la matrice d'alignement final. Cette matrice d'alignement final permet alors d'identifier l'alignement optimal entre la séquence et la structure à l'aide de l'algorithme de NW. Les calculs de score utilisés pour les calculs de la matrice de bas niveau ont été réalisé à l'aide du potentiel DOPE. Les pénalités de gaps ont été fixés à 0.
L'outil DSSP a été utilisé afin d'assigner la structure secondaire associée à une structure.
L'ensemble du code a été écrit en python et les librairies NumPY et BioPython ont été utilisées.

Résultats

Afin de déterminer les performances du programme, celui a été tester sur des séquences et des structures de la famille des globines, qui sont des protéines impliquées dans le transport de l'oxygène. La superfamille des globines est constituée de nombreuses protéines présentant une remarquable diversité de séquences, tout en conservant un repliement commun. La myoglobine est une protéine faisant partie de la famille des globines et est impliquée dans le stockage de l'oxygène dans les muscles. Elle a pour

caractéristiques d'être une protéine constituée principalement d'hélices alpha. Ainsi, afin d'identifier si ce programme est performant, l'enfilage entre 2 myoglobines a été réalisé et comparé au résultat du web server FUGUE (Sequence-structure homology recognition). FUGUE est un programme qui s'appuie sur des tables de correspondances adaptées à l'environnement et des pénalités de gaps qui dépendent de la structure (avec des pénalités différentes pour les hélices alpha, les brins bêta et les boucles). De cette manière, il évalue les scores d'un match entre deux acides aminés et d'insertions/délétions en prenant en compte l'environnement local de chaque acide aminé dans une structure déjà connue. La figure 1 représente les résultats d'alignement entre la structure de la myoglobine de la tortue de mer Caouanne (PDB ID : 1LHT) et la séquence de la myoglobine du *Phoca vitulina,* ayant un pourcentage d'identité de 68.1%. En comparant les résultats de FUGUE et de notre programme, nous remarquons que les résultats obtenus sont identiques. Ainsi, notre programme permet d'avoir des résultats d'alignement correct et identique à ceux de FUGUE dans le cas de myoglobine ayant une forte identité de séquences.

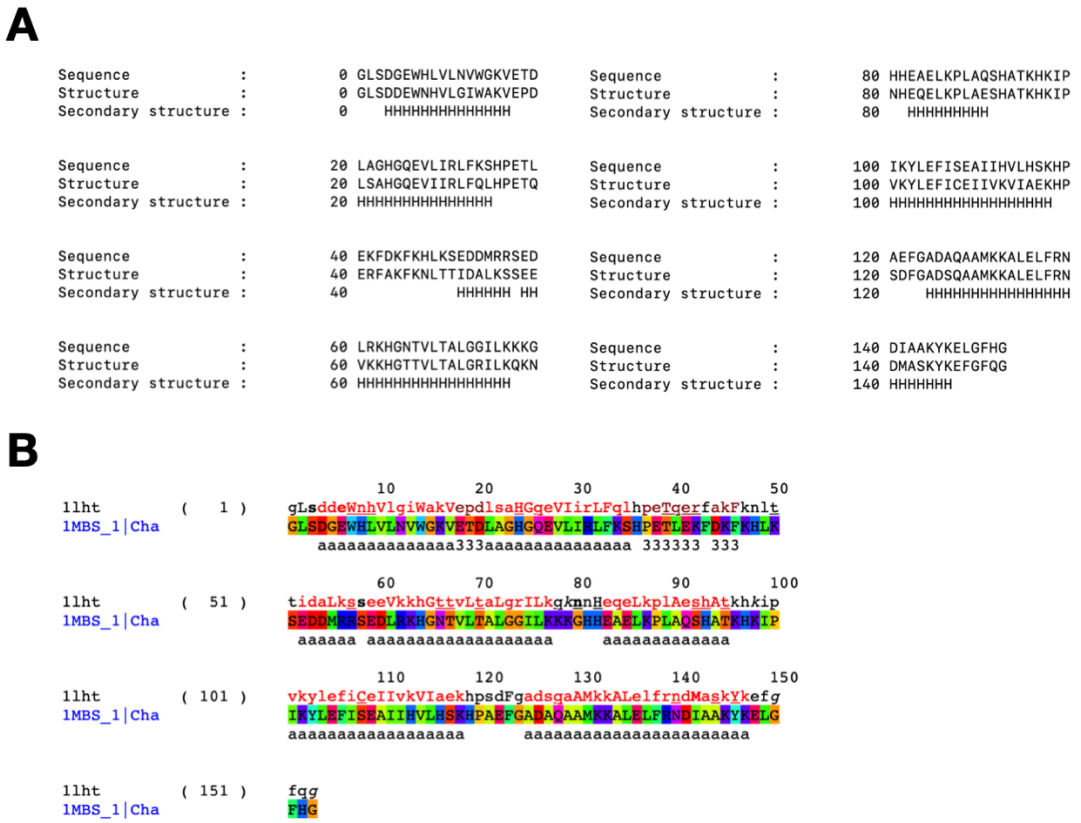

Figure 1 : Alignement entre la stucture de la myoglobine de Caouanne (PDB ID : 1LHT) avec la séquence de la myoglobine du *Phoca vitulina :* **A**. par le biais de notre logiciel. **B**. par le biais du web server FUGUE.

Afin d'identifier si l'alignement optimal obtenu pour ces deux programmes est cohérent avec la réalité, la structure PDB de la myoglobine du *Phoca vitulina* (PDB ID : 1MBS) a été analysée. Afin de vérifier la concordance entre l'alignement optimal obtenu par ces deux programmes et la réalité, nous avons procédé à l'analyse de la structure PDB de la myoglobine de Phoca vitulina (ID PDB : 1MBS). La figure 3 met en évidence que les résidus de la structure 1MBS alignés avec ceux de 1LHT adoptent également une conformation en hélices alpha. En effet, les résidus colorés en bleus sont ceux identifiés par notre programme comme organisés en hélices alpha par le bais du logiciel DSSP. Nos résultats démontrent ainsi la cohérence de notre programme avec les données structurales de la myoglobine *Phoca vitulina*, suggérant une conservation

structurale au sein de cette protéine. Cependant, il est important de noter que la coloration en rose dans la figure 3 indique une légère différence : l'hélice débutant en position 86 apparaît légèrement plus longue. Cette variation est en accord avec les différences observées entre les séquences 1MBS et 1LHT.

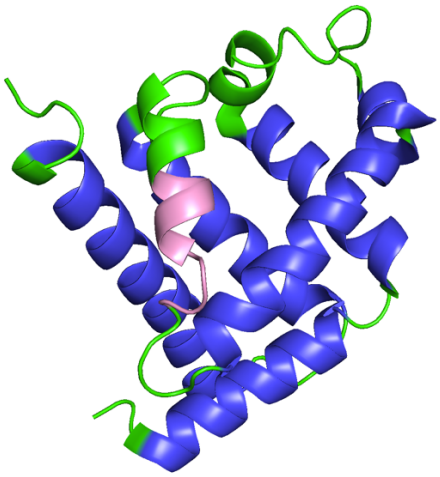

Figure 2 : Structure de la myoglobine du *Phoca vitulina* (Code PDB : 1MBS). En bleu et rose sont représentés les résidus alignés alignés par notre loigiciel avec 1LHT ayant une structuration en hélices alpha.

Étant donné que le programme semble donné des résultats cohérents pour des myoglobines ayant un fort pourcentage d'identité, alors un second test a été réalisé avec cette fois ci 2 protéines ayant un pourcentage d'identité égal à 9.8%. Ainsi, la séquence protéique de la myoglobine de *Caretta caretta* (1LHT) et la structure de leghemoglobin (Code PDB : 1BIN). La figure 3 montre des différences d'alignement entre nos résultats et ceux obtenus par FUGUE.

A

```
Sequence             :         0 GLSDDEWNHVLGIWAKVEPD
Structure            :         0 VAFTEKQDALVSSSFEAF-K
Secondary structure  :         0      HHHHHHHHHHHHHHHHH

Sequence             :        20 LSAHGQEVIIRLFQLHPETQ
Structure            :        20 ANIPQYSVVFYTSILEKAPA
Secondary structure  :        20  HHHHHHHHHHHHHHH   H

Sequence             :        40 ERFAKFKNLTTIDALKSSEE
Structure            :        40 AKDLFSFLANGVDPTNPKLT
Secondary structure  :        40 HHH             HHHHH

Sequence             :        60 VKKHGTTVLTALGRILKQKN
Structure            :        60 GHAEKLFALVR-DSAGQLKA
Secondary structure  :        60 HHHHHHHHHHHHHHHHHHH

Sequence             :        80 NHEQELKPLAESHATKHKIP
Structure            :        80 SGTVVA-DAALGS-VHAQ-K
Secondary structure  :        80      HHHHHHH      HH

Sequence             :       100 VKYLEFICEIIVKVIAEKHP
Structure            :       100 AVTDPQFVVVKEALLKTI-K
Secondary structure  :       100 HHHHHHHHHHHHHHHHHH

Sequence             :       120 SDFGADSQAAMKKALELFRN
Structure            :       120 AAVG-DKWSDELSRAWEV-A
Secondary structure  :       120   HHHHHHHHHHHHHHHHHH

Sequence             :       140 DMASKYKEFGFQG
Structure            :       140 YDELAAAI-KKA-
Secondary structure  :       140 HH
```

B

```
                                10        20        30        40        50
1bin          (   1 )  vaFtekqdalVsssfeaFkanipqySvvFYtSIlekApaAkdlfsfLang
1LHT_1|Cha             -GLSDDEWNHVLGIWAKVEPDLSAHGQEVIIRLFQLHPETQERFAKFKNL
                          aaaaaaaaaaaaaaa aaaaaaaaaaaaaaa    aaaa 333

                                60        70        80        90       100
1bin          (  51 )  vdptnpkLtghAeklFalVrdSAgqLkasgtVvAd--aalGsvhaqkavt
1LHT_1|Cha             TTIDALKSSEEVKKHGTTVLTALGGILKQKNNHEQELKPLAESHATKHKI
                             aaaaaaaaaaaaaaaaaaaaaaa          aaaaaaa

                               110       120       130       140       150
1bin          (  99 )  dpqfvvvk-eALlkTIkaAV----gdkwsdelsrAWevAydelaaaikka
1LHT_1|Cha             PVKYLEFICEIIVKVIAEKHPSDFGADSQAAMKKALELFRNDMASKYKEF
                       aaaaaaa aaaaaaaaaa    333  aaaaaaaaaaaaaaaaaaa

1bin
1LHT_1|Cha             GFQG
```

Figure 3 : Alignement entre la séquence de la myoglobine de *Caretta caretta* (PDB ID : 1LHT) avec la structure de leghemoglobin *:* **A**. par le biais de notre logiciel. **B**. par le biais du web server FUGUE.

Afin de vérifier la concordance entre l'alignement optimal obtenu par ces deux programmes et la réalité, la structure PDB de la myoglobine *Caretta caretta* (ID PDB : 1LHT) a été analysée.

La figure 4A montre que malgré un faible pourcentage d'identité, la structure de 1LHT est proche de celle 1BIN. Ce résultat est attendu étant donné que les globines sont des protéines dont la structure est très conservée. De plus, les résidus de la structure 1LHT (Fig. 4B) alignés avec ceux de 1BIN adoptent également une conformation en hélices alpha, proche de celle de 1BIN (coloration bleu).

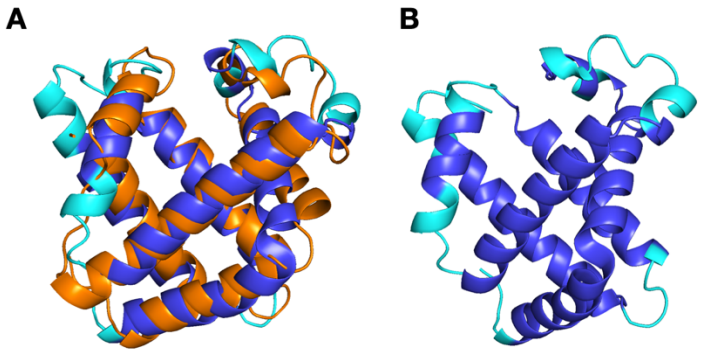

Figure 4 : **A**. Alignement entre la structure 1LHT en bleu et 1BIN en orange. **B**. Structure 1LHT avec en bleu les résidus alignés par notre logiciel avec 1BIN ayant une structuration en hélices alpha.

Conclusion

Par le biais de ce projet, nous avons pu créer un outil capable de réaliser un alignement entre une séquence et une structure. Les résultats obtenus sont cohérents avec ceux obtenus par le biais de FUGUE. Néanmoins, la programme de double programmation dynamique, tel qu'il est décrit dans l'article de David Jones, s'avère trop lent pour être efficace dans le contexte de la reconnaissance des repliements protéiques. Il existe néanmoins des programmes comme FUGUE capable de réaliser de l'enfilage rapidement et de façon précis, contrairement à notre logiciel qui est une version simplifiée de l'outil THREADER.