AVIGNON UNIVERSITÉ

# PRÉDIRE LA STRUCTURE SECONDAIRE D'UNE PROTÉINE EN UTILISANT UN RÉSEAU DE NEURONES - *Application IA* -

**Responsable : Mr Yannick Estève**

Etudiante : Bensafi Sarra

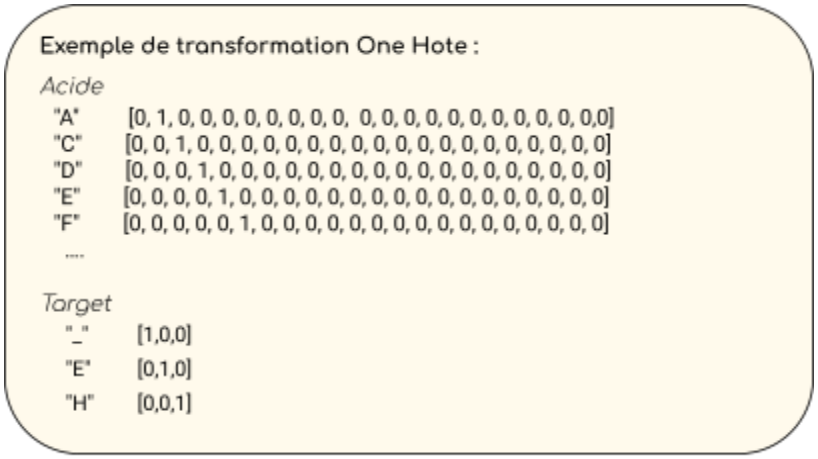
Année : 2021/2022

## 1. Introduction

La prédiction de la structure secondaire d'une protéine est un problème dont l'entrée est une séquence d'acides aminés.

## 2. Données

c'est la partie la plus importante de ce travail , deux approches ont été testées.
Avant de détailler les deux approches, il faut spécifier que les données ont été transformées en One hot. Par exemple :




- ☐ La première approche consiste à entrer la séquence complète de chaque protéine.
  Pour cette partie toutes les séquences devaient avoir la même taille (model GRU) pour cela un remplissage(padding) était nécessaire, en faisant le maximum de longueur pour toutes les protéines, j'ai

remarqué que dans le jeu d'entraînement, la plus longue séquence contient 498 valeurs, le remplissage a été fait avec des zéros.
Ici les acides aminés ont été transformés en one hot avec la fonction prédéfinie de tensorflow.

- ☐ La deuxième approche consiste à utiliser une fenêtre de cinq acides aminés pour prédire celle du milieu, c'est-à-dire en disposant de l'information du milieu, chaque séquence de 5 longueurs sera extraite comme un exemple avec une étiquette.
  Le padding à été utilisé pour les première et dernière acides. Les acides aminés ont été transformés en one hot avec un dictionnaire c'est à dire manuellement.

## 3. Models

*Premier modèle :*
J'ai choisi le meilleur optimiseur et la meilleure fonction de perte en fonction des résultats des nombreux tests que j'ai effectués.
**Optimiseur** : Adam qui était généralement meilleur que SGD.
**Fonction de coût** : MSELoss

**Recurrent Neural Network (GRU)** : j'utilise un GRU à deux couches ce qui signifierait empiler deux GRU ensemble pour former un GRU empilé , le second GRU prenant les sorties du premier GRU et calculant les résultats finaux. Avec input_size= 21 car on a 21 type d'acide aminé (features) à prédir. Le nombre d'entités dans l'État caché h est de 16.

*Deuxième Modèle :*
**Multiple Neural Network (MLP)** :
Pour le MLP, j'ai créé 3 couches linéaires car ce n'est pas un Dataset très large, si on ajoute plus de 4 on aura du overfitting.

- Première couche = 22*5 input ,puis 255 paramètres
- Deuxième couche = 255 , 128 paramètres
- Troisième couche = 128 , 64 paramètres

Toutes les couches sont appliquées avec l'activation ReLU, sauf la dernière couche softmax.
Le Dropout à 0 car en sa présence les résultats n'étaient pas les plus bons.

## 4. Résulatat

| Model | Accuracy | Loss |
|---|---|---|
| MLP no validation | 59.5% | (Entropy) 0.943030 |
| MLP with Validation | 61.0% | (Entropy) 0.924385 |
| GRU Bidirectionnelle(20 epoch) | 63.3% | (MSE)0.060319 |

Les MLP sont limités pour ce genre de données en séquence car ils n'ont pas assez d'information de voisinage et de position et de son contexte.

Les meilleurs résultats obtenus avec le GRU sont dus au fait que ce modèle est récurrent grâce à sa mémoire interne, il se souvient d'éléments importants sur les données qu'il a reçues et permet d'avoir des informations sur ses voisins, et du fait que ce dernier utilise moins de paramètres que son homologue lstm ce qui le rend plus rapide.

Les résultats obtenus avec l'ensemble de validation sont médiocres, ce qui est tout à fait normal puisque les données d'entraînement ont déjà été réduites de 20%  et de base elles sont déjà peu nombreuses.

## 5. Le code(lien)

J'ai essayé de faire un notebook et de commenter chaque fonction pour éviter d'encombrer le rapport avec les captures d'écran.

Code Gru bidirectionnelle  :
https://colab.research.google.com/drive/1yPTAgUUpoMmYlGYV82Zm9feRB49FlAcv?usp=sharing

Code MlP - 5 séquences :
Avec ensemble de  validation :
https://colab.research.google.com/drive/1RqyvGhFx2ANT7Zxiu1bhpyKraqEcpV7C?usp=sharing

Sans ensemble de validation :
https://colab.research.google.com/drive/1R4x-kfRSB8ebpyQ5bEijYNEwcowgFoyU?usp=sharing
Dans ce code j'ai seulement ajouté l'ensemble de validation, sinon c'est exactement le même code que celui d'avant.