Projet 7 : Développez une preuve de concept

Github associé au projet :
https://github.com/MathieuxTony/Projet-7

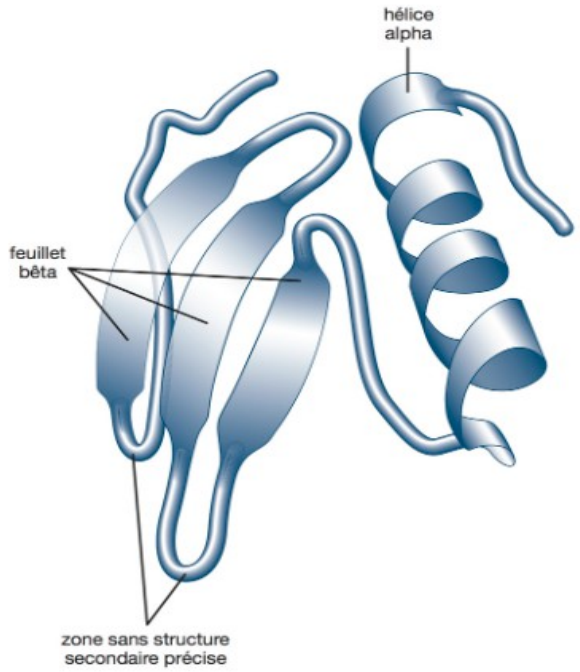
Source principale :
https://arxiv.org/pdf/2208.11248.pdf

1) Le sujet

Les protéines

Les protéines sont des molécules composées d'une succession de molécules plus petites : les acides aminés. La suite d'acides aminés définissant une protéine est la *structure primaire* de la protéine. Il y a 20 acides aminés standards différents et les protéines se composent en générale de quelques centaines d'acides aminées.

La *structure secondaire* décrit le repliement local de la chaîne principale d'une protéine. Il existe trois principales catégories de structures secondaires : les hélices, les feuillets et les coudes.



Objectif

Prédire la position des hélices d'une protéine à partir de sa structure principale, à l'aide de réseaux de neurones.

Prédictions à partir d'une séquence d'acides aminés

Obtenir la structure primaire d'une protéine est relativement simple et peu coûteux, à la différence de la structure secondaire ou d'autres propriétés. Il est donc intéressant de disposer d'algorithmes permettant de déterminer avec précision les propriétés des protéines à partir de leurs séquences d'acides aminés. Il y a eu des avancées majeures dans ces prédictions. Par exemple FoldX permet de prédire l'influence des mutations sur la stabilité, et AlphaFold permet de prédire la structure 3D des protéines (AlphaFold est bien basé sur du machine learning, mais FoldX se base sur des principes physiques).

2) Les données

Source

Il s'agit d'une partie des données se trouvant dans le dossier « data/humans » du github ci-dessous :

https://github.com/malhotra-sidharth/protein-structure-prediction

Format

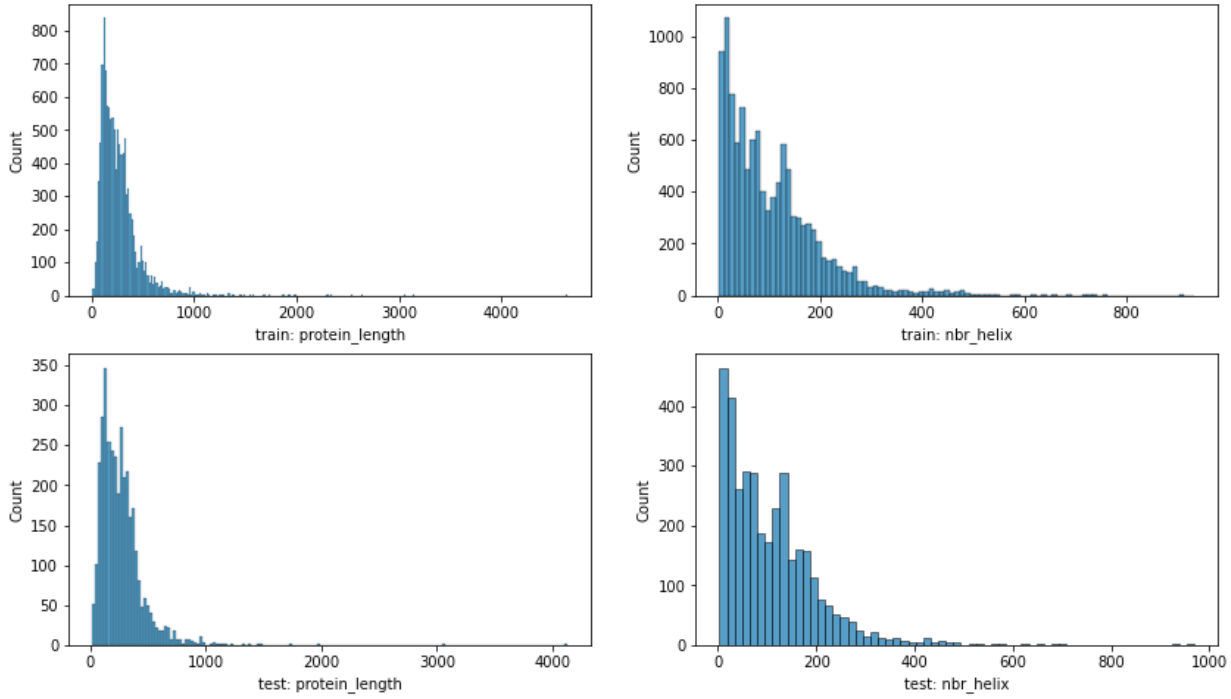
Il y a deux fichiers csv par protéine. Chaque fichier contient une ligne pour chaque acides aminés de la séquence de la protéine, dans l'ordre de la séquence. Dans l'un des fichier les acides aminés sont notés par une chaîne de caractères (non encodé), et d'en l'autre ils sont encodés via un one_hot_encoding. Dans les deux cas il y a aussi une variable binaire indiquant s'il y a une hélice au niveau de l'acide aminé. Il y a un dossier pour les données d'entraînement et un autre pour les données de test.

3) Analyse exploratoire

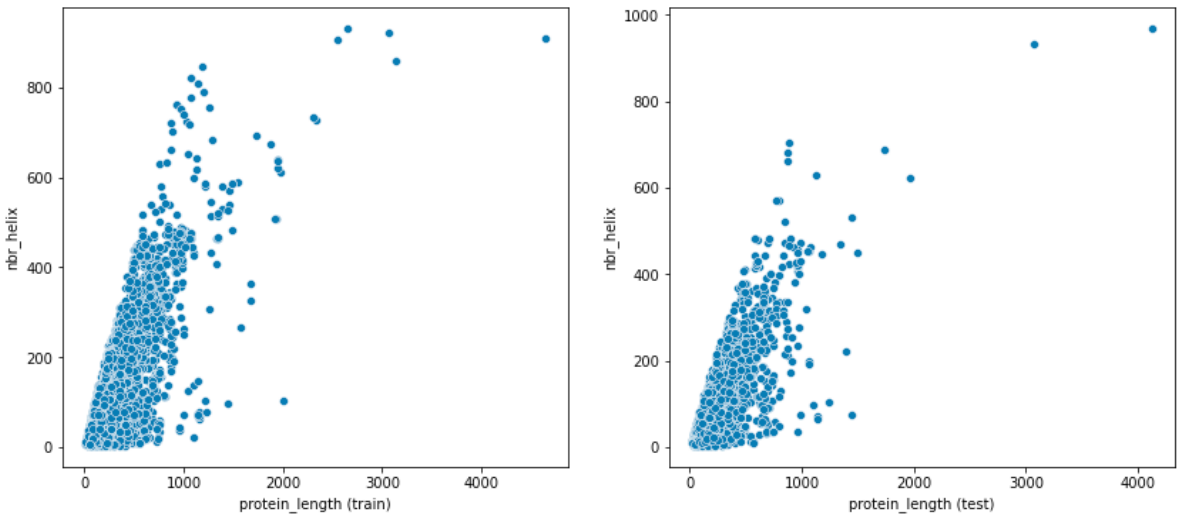
L'analyse exploratoire va surtout nous permettre de vérifier que les données de test ne sont pas trop différentes des données d'entraînement.

En l'état il n'y a pas de variable numérique quantitative à analyser, mais on peut en construire quelques-unes.

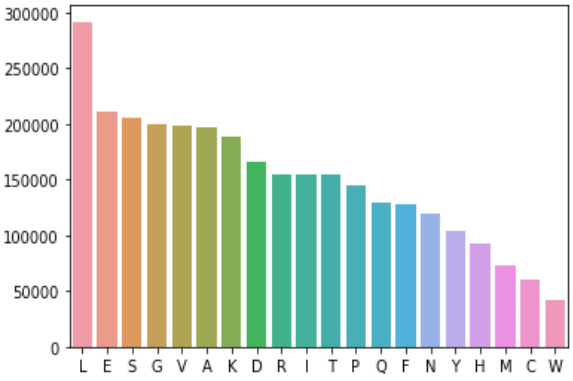
Longueur des séquences et nombre d'hélices



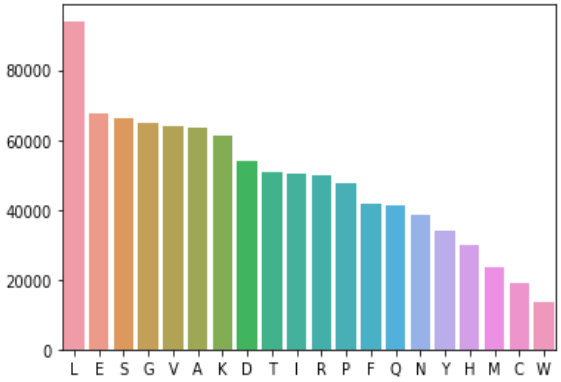
Les histogrammes sont très similaires entre les données d'entraînement et de test. Il en est de même des tracés du nombre d'hélices en fonction de la longueur :



Quantité d'acides aminés



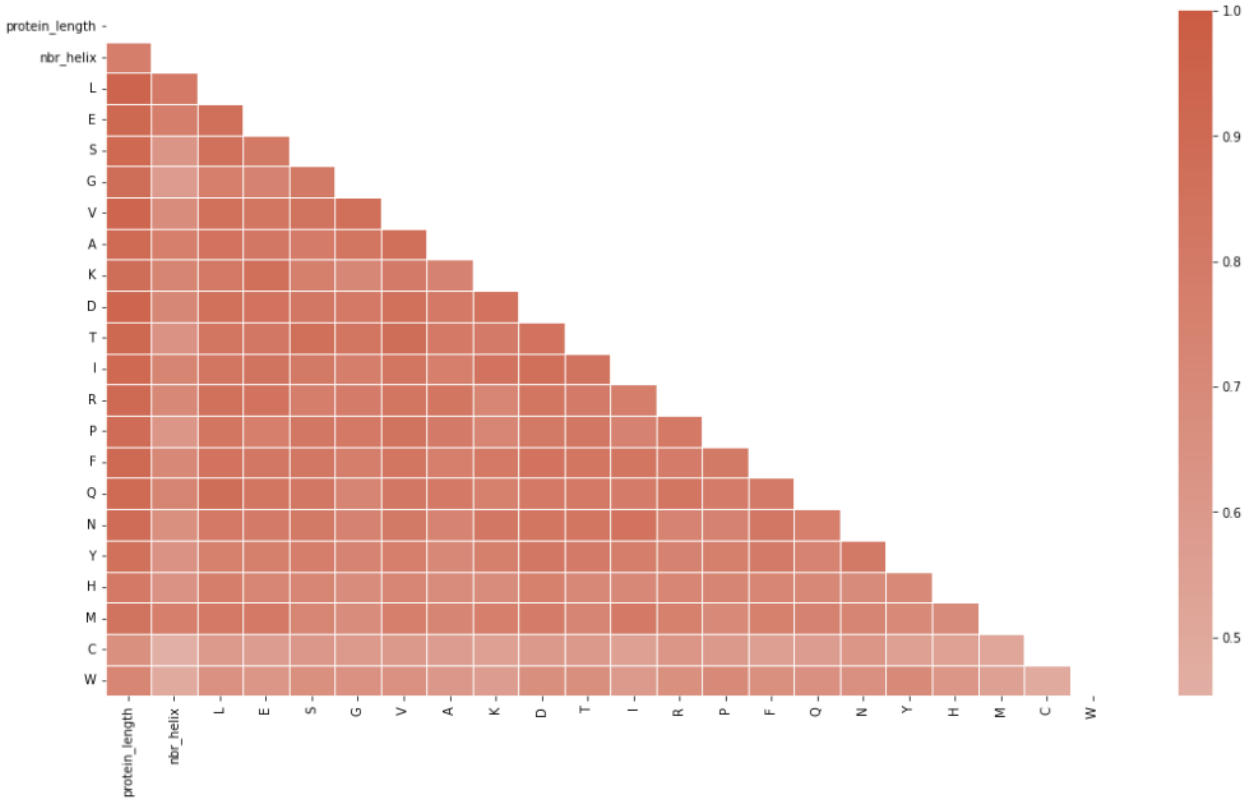
A travers tout le jeu d'entraînement



A travers tout le jeu de test

Là aussi on voit que les données ont été bien répartis.

On peut se demander s'il n'y aurait pas d'acides aminés plus propices aux hélices que d'autres. La matrice de corrélation ci-dessous ne montre pas de corrélation significative.

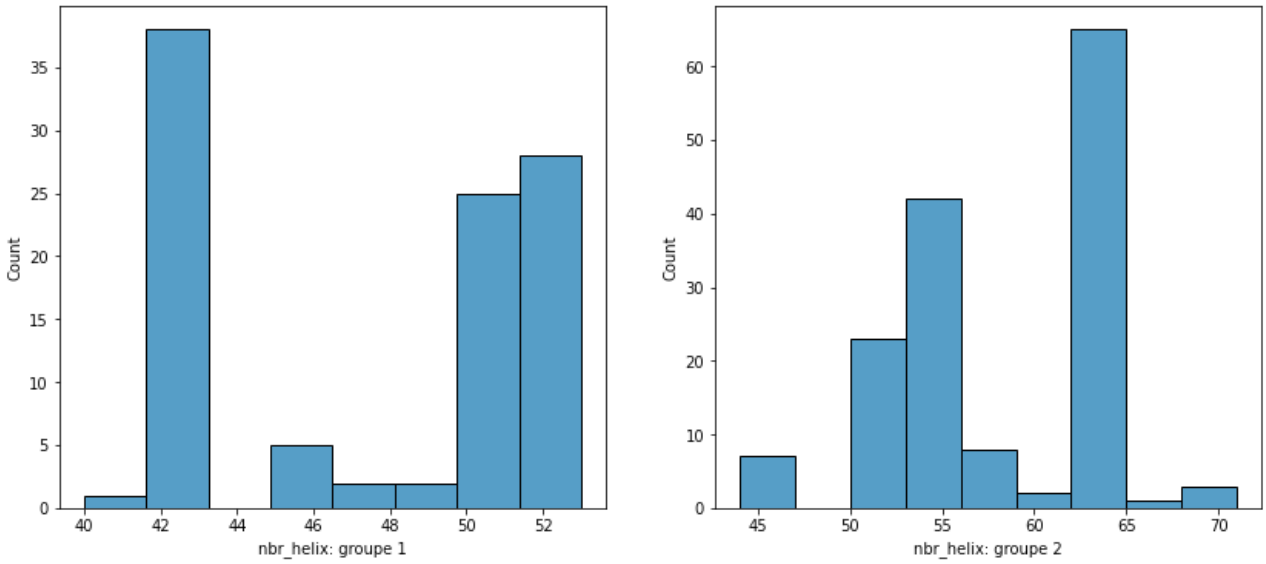


Protéines et mutation

On peut rechercher des séquences de protéines très proches afin de voir si le nombre d'hélices peut fortement varier bien que les structures primaires soient très proches. Pour cela on mesure la distance de Levenshtein entre les séquences. On trouve notamment deux groupes de plus de 100 protéines dont les structures primaires sont très proches.

	protein_sequence	protein_length	nbr_helix	dataset
1	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...	260	42.0	train
23	SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...	259	49.0	train
165	SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...	259	42.0	train
248	SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...	259	42.0	train
249	SHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...	259	42.0	train
...	...	...	...	...
14504	MGMSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSL...	262	50.0	test
14646	AHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPL...	259	53.0	test
14701	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...	260	50.0	test
14710	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...	260	53.0	test
14916	MSHHWGYGKHNGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKP...	260	42.0	test

Un groupe de 101 protéines de structures primaires très similaires



Histogramme du nombre d'hélices au sein de deux groupes de protéines similaires

4) Baseline

Que ce soit avec la baseline ou les réseaux de neurones, nos modèles nous donnerons une probabilité d'avoir une hélice au niveau d'un acide aminé au sein d'une séquence.

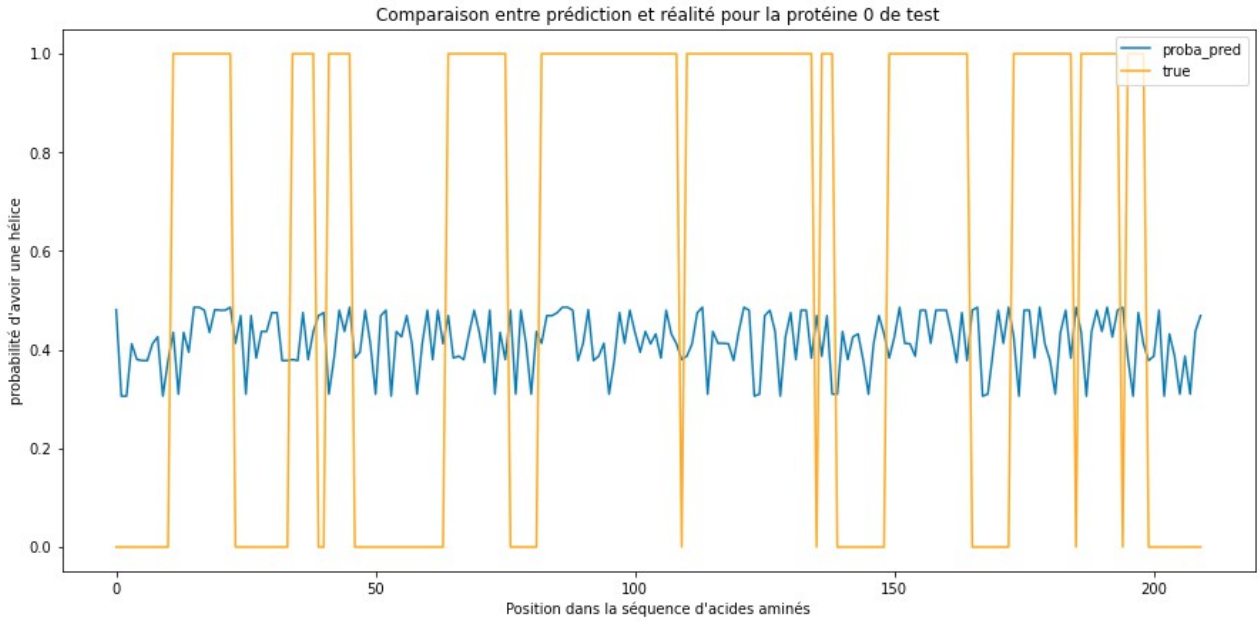
Dans le cas de la baseline, la probabilité d'avoir une hélice au niveau d'un acide aminé donné est égale à la proportion de cet acide aminé, dans le jeu d'entraînement, présentant une hélice. Comme cette approche ne nécessite pas beaucoup de temps de calcul ni d'espace mémoire, on peut utiliser toutes les données d'entraînement pour l'entraînement et de test pour le test.

ALA	0,49	LEU	0,48
ARG	0,44	LYS	0,44
ASN	0,39	MET	0,48
ASP	0,41	PHE	0,43
CYS	0,4	PRO	0,31
GLN	0,47	SER	0,38
GLU	0,48	THR	0,38
GLY	0,31	TRP	0,44
HIS	0,37	TYR	0,41
ILE	0,43	VAL	0,38

Fréquences observées sur les données d'entraînement

L'erreur moyenne en valeur absolue (mae) sur le jeu de test est de 0,477. Ce sera notre performance de référence. Bien sure, ce modèle n'est pas satisfaisant pour une raison évidente : on ne peut pas passer des probabilités à une décision concernant « Oui ou non y-a-t-il une hélice ? » avec un seuil puisque cela revient à décider seulement en fonction de l'acide aminé présent à un endroit donné.

On peut regarder ce que donne la baseline sur une protéine.



Prédiction de la baseline sur une protéine

5) Réseau de neurones

Les vecteurs en entrée

Nous allons utiliser une « fenêtre glissante » de longueur 10, c'est-à-dire que chaque protéine fournira beaucoup plus qu'un vecteur en entrée (précisément, la longueur de sa séquence moins 9) : ces vecteurs seront constitués de 10 acides aminés consécutifs, encodés par one_hot_encoding, parcourant toutes la séquence de la protéine. Pour chacun de ces vecteurs il y aura un vecteur à prédire, constitué de 10 valeurs 0 ou 1 correspondant aux hélices.

Ainsi une protéine constituée de 600 acides aminés donnera un dataframe de 591 lignes et 210 colonnes (200 pour les acides aminés encodés et 10 pour les hélices à prédire) et on se retrouve vite limité par l'espace mémoire. Nous utiliserons 5000 protéines pour l'entraînement et 1000 pour le test.

En sortie

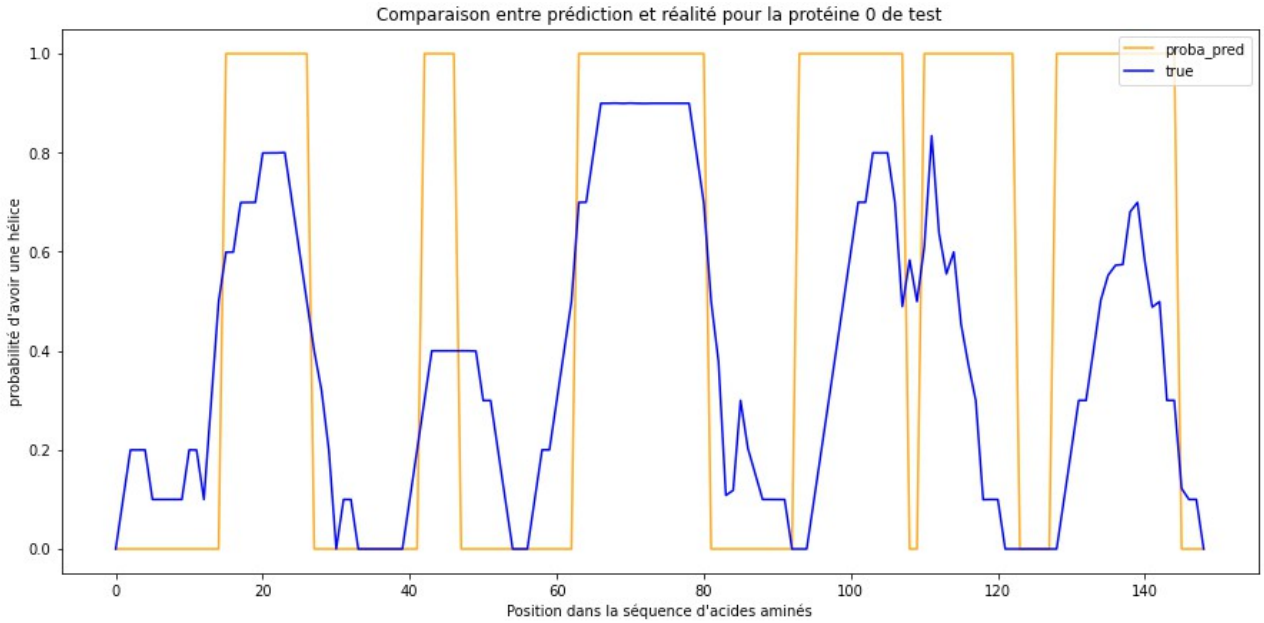
La plupart des acides aminés seront contenus dans plusieurs fenêtres et on aura alors plusieurs valeurs de probabilité prédites. Pour chaque acide aminé d'une séquence, on fera une moyenne des probabilités obtenues sur les fenêtres qui le contiennent.

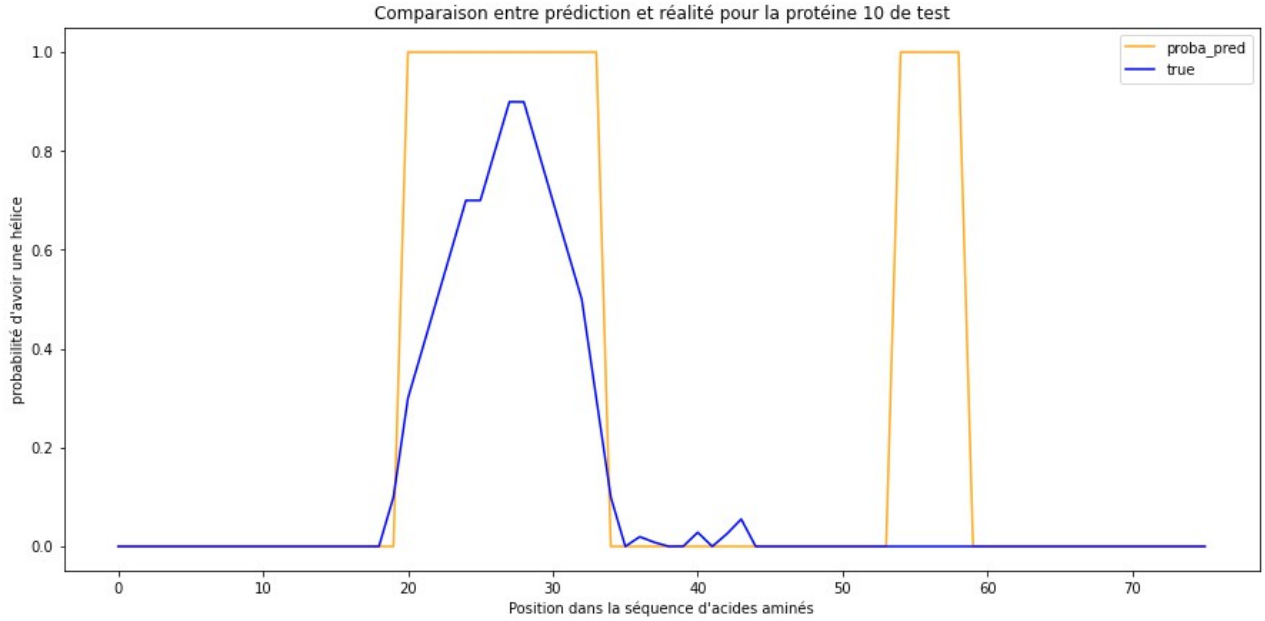
Le modèle

Le modèle se compose de trois couches de neurones successives complètement connectées contenant respectivement 200, 40 et 10 neurones. Toutes les fonctions d'activation seront 'relu'.

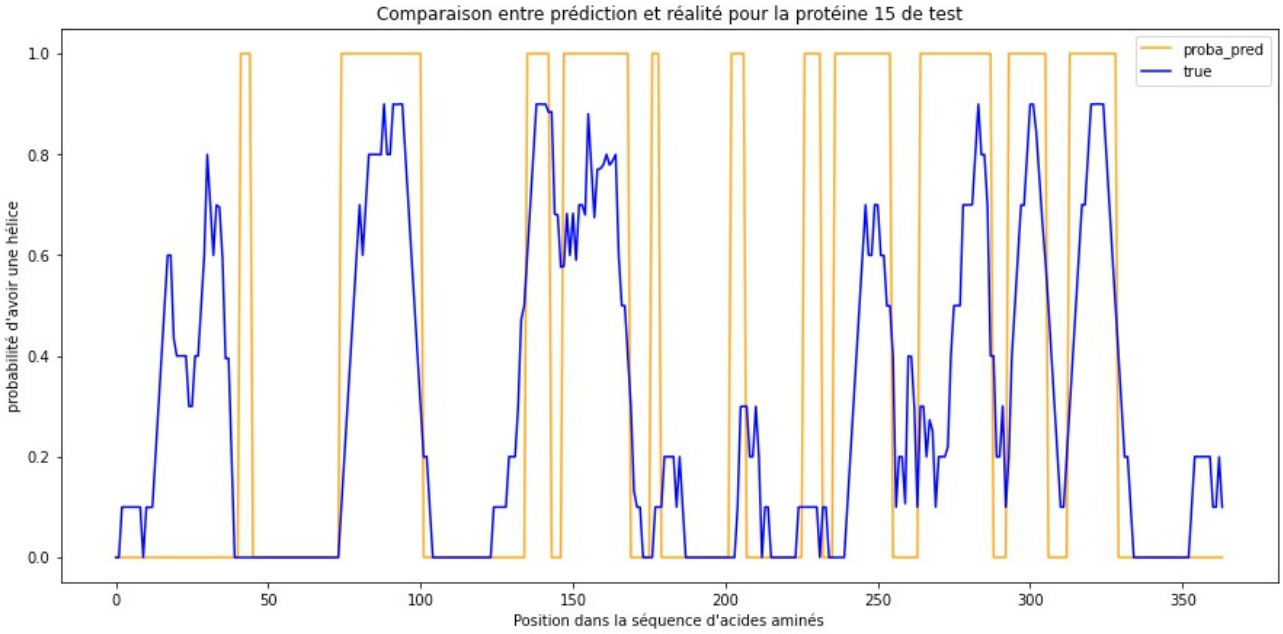
Résultats

On obtient une mae d'environ 0,29 ce qui est significativement meilleur que la baseline. Il faudrait encore convertir les probabilités en décisions binaires, par exemple avec un seuil de 0,5, mais nous allons conserver les probabilités pour faire des représentations graphiques sur des exemples afin de voir les « tendances ». Voici ci-dessous quelques exemples de prédictions comparées à la réalité.





Ci-dessus une hélice à été clairement manquée.



Quand il y a beaucoup d'hélices la prédiction oscille plus, on voit les tendances mais on va manquer de précision en passant à une prédiction binaire.

Bibliographie

https://arxiv.org/pdf/2208.11248.pdf

https://github.com/malhotra-sidharth/protein-structure-prediction

https://en.wikipedia.org/wiki/Amino_acid

https://medium.com/predict/deepmind-ai-predicts-protein-structure-48cba338b84b

https://medium.com/analytics-vidhya/stock-price-prediction-single-neural-network-with-tensorflow-75b64af74ed6

https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39