Projet Deep Learning :
Classification of ligand-binding pockets in proteins

Ce projet est basé sur l'article: Pu L., Govindaraj R.G., Lemoine J.M., Wu H.C. & Brylinski M. (2018). DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput Biol, 15(2)*.

I. Introduction

Il existe une multitude de type de protéine différentes. L'être humain en fabriquerait plus 100 000 sortes différentes parmi lesquelles ont retrouve des stéroïdes, des métalloprotéines, ... Il est fréquent que les protéines se lient avec des ligand. Un ligand est une molécule qui se lie de manière réversible à une macromolécule ciblée comme un protéine. Il peut jouer un rôle dans la stabilisation de la molécule, jouer un rôle de catalyseur ou encore transmettre un signal. Les protéines ont des formes bien spécifique que l'on pourrait comparer à des "poches" et le ligand vient si encastrer. Comprendre comme un ligand interagit avec une protéine est un enjeux scientifique majeur. En effet, il est important de comprendre ces interactions afin de pouvoir étudier les fonctions moléculaires de certaines protéines mais également pour comprendre les liens évolutifs entre elles. D'un points de vu technologie, l'étude des ces interactions permet également de développer des enzymes et des substrats de synthèse utiles dans le développement de nouvelles médications. Cependant, malgré la technologie grandissante, il reste compliqué d'étudier avec précision ces interactions. Les protéines peuvent parfois présenter de grandes similarité structurales mais pourtant correspondre à des ligands différents. La vitesse de production de données étant en augmentation il devient important de pouvoir discriminer une protéine en fonction de son ligand.

Dans l'article sur lequel ce travail est basé et inspiré ce travail, les chercheurs ont tout d'abord commencé par modélisé la structure moléculaires de différentes protéines. Pour ce faire, ils ont modélisé une sphère de rayon 15 Å de voxels espacés de 1 Å autour de la protéine. Les voxels sont au fur et à mesure éliminés de la sphère: en premier les voxels superposant la protéine; Ensuite les voxels trop éloignés de la protéines. A chaque voxels restant est attribué un potentiel d'interaction électrostatique. Il existe 14 interactions possibles. La poche ainsi modélisé est alors placé dans un cube de 32x32x32. Pour ce projet, les données qui sont utilisés ont déjà été mises sous cette forme.

On dispose ici d'un jeu de données contenant 596 poche de hèmes, 1553 poche de nucléotides, 69 poches de stéroïdes et 1946 poches contrôles. Les hèmes sont des cofacteurs contenant un atome de métal, souvent du fer, servant à accueillir un gaz diatomique (par exemple du dioxygène O2). Parmi les protéines contenant des hèmes on retrouve notamment les hémoglobines ou encore le cytochrome C par exemple. Le groupe contrôle est constitué de protéines et de poches qui ne sont pas dans les groupes hèmes, nucléotides ou stéroïdes. Pour chaque poche il y a une protéine associée.

Le but de se projet est donc de créer un réseau de neurones capables de discriminer la nature d'une partie de la structure d'une protéines entre hème, nucléotide, stéroïde ou autre en fonction des potentiels d'interaction électrostatiques entre la protéine et son ligand.

II. Matériels et méthodes

Tout les fichiers et documents sont disponibles sur GitHub (https://github.com/nicolassilva/DeepLearningProject_DeepDrig3D). Ce projet a été réalisé sous un environnement Conda avec Python 3.7. Sur cet environnement ont été également installés Numpy, Random, Keras and Tensorflow. Sur le git, est disponible un fichier *.yml* avec lequel la création d'un environnement Conda identique à celui dans lequel le projet a été réalisé est possible.

La première étape du programme est de créer un jeu d'apprentissage pour le réseau. Les données de voxels sont bien données sous formats binaire ce qui facilite l'importation des data par le programme python. Cependant, il est nécessaire ici de créer un array au format binaire qui permet d'indiquer a quel groupe (hème, nucléotide, stéroïde ou contrôle) appartient la poche d'intérêt. Pour la partie apprentissage, il est important de réduire le jeu de données car l'analyse de la totalité du jeu de données surpasse la mémoire d'un ordinateur portable classique. Comme il y a beaucoup moins de stéroïdes que des autres classes j'ai dupliqué la totalité des stéroïdes N fois. N correspondant à la centaine des autres classes sélectionnées (par exemple pour 300 hèmes, nucléotides et contrôles, alors N=3). Cela évite que les stéroïdes ne se retrouvent noyés dans la masse de données des autres classes. Il est important de noté que poches sélectionnées doivent au mieux se faire au hasard. Ici la sélection se fait dans l'ordre pour l'apprentissage et par la fin de la liste pour le test afin de ne pas prendre les même poches pour l'apprentissage et pour le test. Les listes ainsi créées sont ensuite mélangées afin de ne pas avoir toutes les poches d'un mêmes groupe à la suite pour l'apprentissage. On définit ensuite la dimension du canal en "Channel first". On peut définir un canal de deux manières en prenant comme exemple ici les données: la première manière est de donné d'abord les dimensions du cube autour de la poche et ensuite le type d'interaction obtenant ainsi des données aux dimensions 32x32x32x14 ("Channel last"); la seconde manière est de donné d'abord le type d'interaction et ensuite les dimensions du cube autour de la poche obtenant ainsi des données aux dimensions 14x32x32x32 ("Channel first").

Plusieurs réseaux on été testés ici:

- Le premier réseau (modèle A) est composé d'une couche de convolution 3D constitué de 16 neurones et d'une taille de kernel de 2x2x2. Il y a ensuite un drop_out de 20% suivie d'un maxPooling de 2x2x2 et d'encore un drop_out de 20%.
- Le second réseau (modèle B) est composé de deux couches de convolution 3D constitué de 4 neurones et d'une taille de kernel de 2x2x2 pour la première et de 16 neurones et d'un kernel_size de 2x2x2 pour la seconde. Il y a ensuite un drop_out de 20% suivie d'un maxPooling de 2x2x2 et d'encore un drop_out de 20%.
- Le troisième réseau (modèle C) est composé de deux couches de convolution 3D constitué de 16 neurones et d'une taille de kernel de 2x2x2 chacune. Il y a ensuite un drop_out de 20% suivie d'un maxPooling de 2x2x2 et d'encore un drop_out de 20%.

Tous les modèles implémentés finissent avec une couche flatten et une couche dense de 4 neurones avec une activation softmax. Ils sont implémentés avec un batch size de 16, un

validation split de 30% et un nombre d'epochs de 300. Afin d'éviter le sur-apprentissage les modèles ont un earlyStopping avec une patience de 10.

L'évaluation des modèles a été effectuée sur un jeu de données comprenant 300 poches de chaque classes sauf pour les stéroïdes qui ne contient que 69 poches. Les poches sélectionnées sont différentes de celles utilisées pour l'apprentissage. Ensuite des courbes ROC (Receiver Operating Characteristic) ont été créées pour chaque classes dans chacun des modèles implémentés.

III. Résultats

Tableau 1: Perte et Précision des différents modèles.

	Model_1conv3D_16 (A)	Model_2conv3D_4_16 (B)	Model_2conv3D_16_16 (C)
Perte	11.13	0.53	0.32
Précision	0.31	0.80	0.90

L'analyses des performances globales montre clairement des différences entres les modèles. En effet, on peut voir que les modèles B et C présentent une précision nettement supérieur au modèle A de l'ordre de 50 à 60% de plus (Tableau 1). Cependant, comme les analyses menées ici sont du type multi-classes, il est nécessaire de s'attarder sur les performances de discrimination classes par classes du modèle.

Les courbes ROC permettent de regarder la sensibilité et la spécificité des modèles afin de pourvoir déterminer de sa fiabilité à prédire la bonne classe de poches et à éviter les faux positifs. L'analyses des courbes ROC confirme que le modèle A est le modèle le moins efficace des trois. De plus, ces courbes montrent ici que le modèle C est bien lui le plus efficace. On peut voir sur les courbes ROC du modèle C que celui-ci discrimine les différentes poches avec très peu de faux positifs et ce pour n'importe quelle des quatre classes. Les modèles B et C atteignent même 100% de précision quand à la prédiction de poches de stéroïdes.

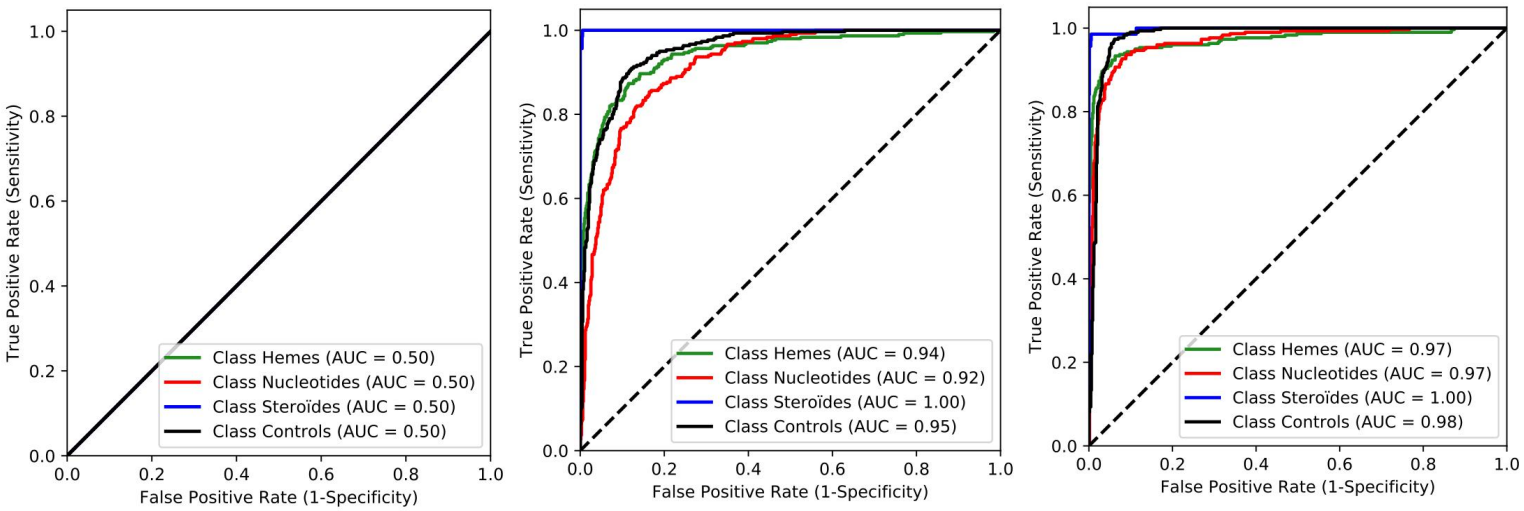

Figure 1: Courbes ROC pour les trois modèles implémentés ici. De gauche à droite les courbes ROC des modèles A, B et C. Dans la légende le terme AUC correspond à l'aire sous la courbe.

IV. Conclusion

Le modèle C implémenté ici est relativement efficace pour discriminer les poches Hèmes, Nucléotides et Stéroïdes de celles qui n'en sont pas (les poches Contrôles). En effet Il présente une précision de classification supérieure à 97% pour les Hèmes, les Nucléotides et les Contrôles, et même une précision de 100% pour les stéroïdes. Cependant, la précision pour les stéroïdes est à prendre avec légèreté. Les stéroïdes ne sont présent ici qu'en petit nombre et on été mis plusieurs fois pour qu'il ne soient pas noyé dans la masse de données de poches des autres classes. De plus, comme le nombre de poches Stéroïdes n'est que de 69, les poches utilisées pour la partie test du modèle sont les même que celles utilisées dans l'apprentissage. De ce fait, la précision de 100% dans la discrimination des Stéroïdes vis à vis des autres classes s'en retrouve biaisée.

Les modèles testées ici sont légèrement différents de ceux de l'article. En effet, le nombre de neurones impliqués dans la double convolution 3D implémenté ici est plus faible (2 fois 16 neurones contre 2 fois 64 neurones dans l'article). Cependant un tel réseau nécessite des machines avec de plus grande capacité. De plus au vu des résultats et de la précision de discrimination des classes autres que les Stéroïdes, un réseau avec moins de neurones ne présente que légèrement moins de performance que le réseau implémenté dans l'article.

Après les différents test effectués, il y a cependant des conditions nécessaires pour que le réseau de neurones apprenne. En effet, dans les essaies que j'ai pu effectué, j'ai remarqué que pour des données comme celles-ci, une seule couche de convolution 3D était inefficace et que pour que le réseau apprenne un peu il était nécessaire de double cette couche. De plus, le mélange des données avant l'apprentissage est également important. En effet, lorsque les données ne sont pas mélangés et présentées dans l'ordre au réseau celui-ci présente de grandes difficultés à apprendre.