**Rapport Deep Learning**

Julie GORSE BOGOIN

M2 BIB 2019-2020

Lien GitHub : https://github.com/juugeebee/Projet\_DL\_M2

**INTRODUCTION**

1. **Le Deep learning (apprentissage automatisé)**

Le deep learning une classe d’algorithmes d’apprentissage automatique. Dans le cadre d’un apprentissage supervisé, l’objectif est de classer des données en fonction de caractéristiques choisies et communes. Les réseaux de neurones en sont un des outils et ils permettent à l’information de traverser plusieurs couches d’apprentissage afin que la prédiction soit la plus juste possible. Pour cela le réseau doit « s’entrainer » sur un jeu dit d’apprentissage composé d’un échantillon de données dont on connait la classe. Si le modèle est correctement paramétré, il présentera des capacités de généralisation sur des données qu’il n’a jamais rencontrées. Un réseau de neurones à convolution est un type de réseau de neurones artificiels dans lequel le motif de connexion entre les neurones est inspiré par le cortex visuel des animaux. Les neurones consistent en un empilage multicouche de perceptrons dont le but est de prétraiter de petites quantités d'informations. Ils ont de larges applications dans la reconnaissance d'image et vidéo, et le traitement du langage.

1. **L’outils DeepDrug3D**

Dans leur papier de février 2019 publié dans Computational Biology, *Brylinski et al.* décrivent un outils de détection et de classification de sites de fixation médicamenteux utilisant un réseau de neurones à convolution : DeepDrug3D. Il est open source et les scripts sont accessibles sur GitHub. Le challenge dans cette problématique est de développer un modèle qui permettra une détection fine pour des molécules ayant des structures biochimiques proches. Les organisations structurales des différents sites de fixation testées ont été représentées par les auteurs dans des voxels (grilles d’interaction tridimensionnelles). Ces voxels sont les données d’entrée du réseau de neurones. L’implémentation actuelle de DeepDrung3D est entrainée pour détecter et classifier des sites de fixation de nucléotides et d’hèmes. Elle a aussi la capacité de généraliser sur des sites de fixation de stéroïdes et de peptidases. Il a été démontré que DeepDrug3D faisait une distinction précise entre des poches qui ont des structures spatiales proches.

**METHODES**

1. **Librairies utilisées**

* numpy (Numerical Python) fournit une interface pour stocker et effectuer des opérations sur les données.
* os fournit une manière portable d’utiliser les fonctionnalités dépendantes du système d’exploitation.
* random permet de gérer l’aléatoire dans le script.
* matplotlib constitue un puissant outil pour tracer des graphes afin de visualiser des données.
* tensorflow est une librairie open-source pour le calcul numérique dans laquelle sont utilisées des graphiques de flux de données.
* keras est une librairie de réseaux de neurones.

1. **Jeux de données**

Pour réaliser ce projet, 4 inputs ont été nécessaires :

* 4170 fichiers binaires *\*.npy* (NumPY data file), voxels, représentant les sites de fixation de tous types.
* Une liste de 1553 noms de voxels impliqués dans des interactions avec des nucléotides
* Une liste de 596 noms de voxels impliqués dans des interactions avec des hèmes
* Une liste de 69 noms de voxels impliqués dans des interactions avec des stéroïdes

Il est à noter que les poches stéroïdes sont bien discriminées des poches hèmes par DeepDrug 3D alors que leurs structures spatiales sont très proches.

* Une liste de 1946 noms de voxels control qui ne se fixent ni aux nucléotides, ni aux hèmes (et ni aux stéroïdes par extension).

1. **Description du modèle**

Le tableau ci-dessous décrit couche par couche le réseau de neurones mis en place pour ce projet. Sont détaillés ici, leurs types, leurs formats de sortie et les paramètres utilisés.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Layer (type)** | | **Output Shape** | | **Parameters** |
| conv3d\_1 (Conv3D) | | (None, 14, 32, 32, 32) | | padding='same' |
|  |  |  |  | activation = 'relu' |
| dropout\_1 (Dropout) | | (None, 14, 32, 32, 32) | | 0,2 |
| max\_pooling3d\_1 (MaxPooling3) | | (None, 14, 16, 16, 16) | | padding='same |
| dropout\_2 (Dropout) | | (None, 14, 16, 16, 16) | | 0,4 |
| flatten\_1 (Flatten) | | (None, 57344) | |  |
| dense\_1 (Dense) | | (None, 512) |  | 512 |
| activation\_1 (Activation) | | (None, 512) |  | activation = 'relu' |
| dense\_2 (Dense) | | (None, 3) |  | 3 |
| activation\_2 (Activation) | | (None, 3) |  | activation = 'softmax' |

1. **Méthodes d’évaluation**

Pour évaluer le modèle, des indicateurs sont disponibles. Ils sont calculés en tenant compte de la mauvaise ou bonne association sites / classes.

* Un vrai positif est un site qui est classé dans sa classe.
* Un faux positif est un site pour lequel le modèle a attribué une classe alors qu’il appartenait à la classe contrôle.
* De la même façon, un vrai négatif est un site contrôle pour lequel aucune classe n’a été déterminée.
* A l’inverse un faux négatif serait un site dont on connait la classe mais qui aurait été étiqueté sans classe.

A partir de ces comptages, comme décrit dans la publication de *Brinlinsky et al*, il est possible de calculer :

* ACCuracy
* Positive Predictive Value
* True Negative Rate (Spécificité)
* True Positive Rate (Sensitivité)
* False Positive Rate

Les spécificité et sensibilité permettent de tracer des courbes de ROC (Receiver Operating Characteristic). Elles sont des mesures de la performance d'un classificateur, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en groupes distincts. L

L’aire sous la courbe de ROC est également très discriminante puisque plus elle est proche de 1 et plus les prédictions sont fiables.

**RESULTATS**

* La précision de la prédiction du modèle est de 66% C’est assez faible comparé au 94% obtenu par Deep Drug 3D.
* Le score de prédiction est de 76%, c’est plutôt bon surtout quand on sait que le score de Deep Drug 3D est de 81%.
* Le taux de vrais positifs est de 47%. Cela signifie que la moitié seulement des sites de fixation ont été discriminés correctement. Ce résultat est peu satisfaisant. Deep Drug 3D atteint quant à lui un taux de vrai positif de 91%.
* L’assignation des négatifs est encore plus mauvaise puisque seulement 19% des négatifs sont classés contrôle. Deep Drug 3D atteint un taux de 96% dans ce domaine.

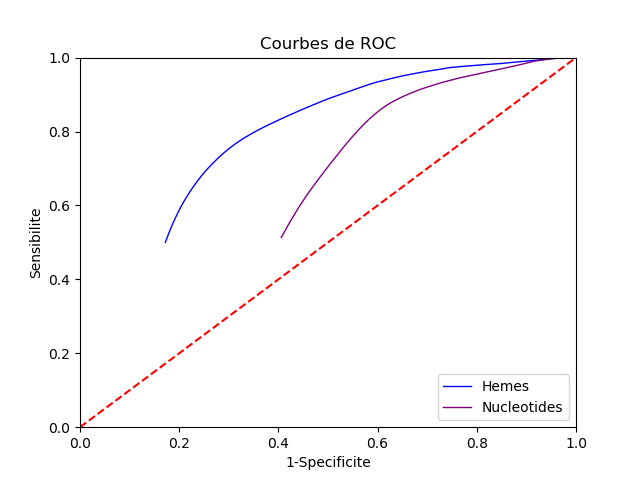


Figure 1 : Courbes de ROC des jeux de données nucléotides (en bleu) et hèmes (en violet)

La figure 1 représente la sensibilité du modèle en fonction de 1-sa spécificité. Plus un modèle est précis et plus les courbes tendent vers le point (0,1). Ici il est notable que les deux courbes ne commencent pas à (0,0). C’est étrange ; l’implémentation de la fonction de calcul des points doit être en cause. D’autre part, il est possible de noter que l’allure de ces deux courbes est assez plate. Cela est tout à fait en corrélation avec les résultats assez mitigés que les méthodes d’évaluation par calculs avaient mis en évidence.

Les aires sous la courbe des deux classes ont été calculées :

* Nucléotides : 0,52
* Hèmes : 0,76

Réciproquement à la courbe de ROC et aux paramètres arithmétiques, ces deux valeurs sont assez loin des performances attendues. En effet, pour être optimum, l’air sous la courbe doit tendre vers 1 indiquant ainsi que le taux des fausses assignations de classes est le plus négligeable possible. Pour mémoire, les valeurs moyennes d’aire sous les courbes obtenues par Deep Drug 3D sont de 0,98 soit approchant la prédiction parfaite.

**CONCLUSION**

Prises indépendamment de DeepDrug3D, les performances du modèle mis en place ne sont pas très satisfaisantes. Il faudrait, pour les optimiser, modifier les différentes couches et leurs paramètres afin de comparer les résultats obtenus. Après comparaison avec DeepDrug3D, la perception est encore plus sévère, ses performances étant excellentes.

Cela étant, pour une première expérience de création de modèle de réseau de neurones à convolution, il semble réconfortant de se dire que d’avoir réussi à produire un code qui fonctionne et qui discrimine correctement plus de la moitié des jeux de données test est déjà une belle réalisation. En disposant de plus de temps et d’un matériel plus performant, il aurait été possible d’optimiser le modèle afin de pouvoir le tester sur le jeu de données stéroïdes, plus difficilement classifiable.

Deep Drug 3D étant en libre accès, si la question scientifique se posait de devoir discriminer des sites de fixation en fonction de la protéine qui s’y liguera, Deep Drug 3D serait une solution absolument pertinente.