République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie d'Oran – Mohamed Boudiaf
Faculté des Mathématiques et Informatique
Département d'informatique

TP 1 Bio-informatique

Classification et conception spécifique de l'amorce pour la détection précise du SRAS-CoV-2 à l'aide du deep learning

Domaine : **Mathématiques – Informatique**
Filière : **Informatique**
Spécialité : **M2 IAA**

Présenté par :

- KEBIRI Issam Dine

- AIT AMRANE Toufik

Supervisé par :

- Mme Drioua.w, department d'informatique - USTO.

Décortication de l'article :

Après avoir lu l'article plusieurs fois, on distingue clairement plusieurs informations, méthodes et résultats à décrire. Mais la principale remarque que j'en déduits de cet article est la puissance de la méthode CNN (convolutional Neural Networks) ainsi qu'au tôt élevé de la fiabilité des résultats. Les principaux points à retenir :

1. Le coronavirus appartient à la famille des Coronaviridae qui affecte les hôtes aviaires et mammifères, y compris les humains. En tant que virus ARN typique, de nouvelles mutations apparaissent à chaque cycle de réplication du coronavirus.
2. Les tests PCR manque de fiabilité pour la détection positif ou négatif du virus à cause de la fréquence de mutation élevée de ce dernier ainsi que la similitude avec d'autres infections respiratoires de la famille du corona virus.
3. La classification traditionnelle se fait à l'aide de techniques de séquençage viral qui est principalement basée sur des méthodes d'alignement telles que BLAST. Ces méthodes ont leur avantages et inconvénients.
4. Compte tenu de l'impact de l'épidémie mondiale, des efforts internationaux ont été déployés pour simplifier l'accès aux données génomiques virales et aux métadonnées par le biais de dépôts internationaux tel que NGDC, NCBI et le GISAID.
5. Le principal travail du CNN est de séparer les coronavirus appartenant à différentes souches, y compris le SRAS-CoV-2. On génère ensuit des séquences représentatives de l'ADNC que le réseau utilise pour classer le SRAS-CoV-2. Après validation des séquences découvertes les résultats montre que les classificateurs traditionnels et simples à évaluer correctement le SRAS-CoV-2 avec une précision remarquable (> 99 %).
6. Quelques-unes des séquences découvertes possèdent également les caractéristiques correctes pour devenir des amorces, plus incroyable le CNN arrive même à générer les principales séquences de différents ensembles d'amorce mis au point par les laboratoires de référence de l'Organisation mondiale de la Santé (OMS) !
7. Le mode de fonctionnement du CNN est assez simple à comprendre : le CNN est composé de 4 couches, une couche convolutionnelle avec 12 filtres (ou pois si on le veut) chacun avec la taille de la fenêtre 21, une

couche entièrement connectée et une couche softmax finale avec 5 unités (5 unités en références des 5 classes de souches de coronavirus) plus un optimiseur Adaptive Momentum.

8. Une fois le CNN prêt on lance la première analyse et on rapport la visualisation des 1250 premier points des échantillons rapporté de la NGDC Référentiel, étants donné les filtres avec leur sorties Boolean (1 ou 0) les échantillons appartenant à différentes classes peuvent déjà être distingués visuellement grâce aux 12 filtres de la couche convolutionnelle, et un filtre se démarque car il semble se concentrer sur quelques points pertinents dans le génome, qui pourraient correspondre à des séquences significatives d'ADNC du SRAS-CoV-2.
9. Apre cette étape il est maintenant possible d'identifier les séquences de 21 points de base qui ont obtenu les valeurs de sortie les plus élevées dans la couche de mise en commun maximale du filtre qui s'est démarquée, et on obtient alors des séquences uniques pour le SRAS-CoV-2.
10. Exemple dans l'article d'une séquence qui se trouve qu'à l'intérieur de la classe du SRAS-CoV-2 : AGG TAA CAA ACC AAC CAA CTT. Encore une information remarquable : le CNN peut identifier les séquences même si elles sont légèrement déplacées dans le génome.
11. Dernière information qui montre la puissance de cette méthode : 99% des séquences de différents ensembles d'amorce utilisées dans les tests RT-PCR SARS-CoV-2 mis au point par les laboratoires de référence de l'Organisation mondiale de la Santé (OMS) ont été trouvé dans cette étude.