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 SARS-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.

Plan :

Figures :

I. Introduction générale :

Le but de ce TP1 est de choisir un article récent (à partir de janvier 2020) de la base de données PubMed et le décortiquer, nous avons voulu choisir un article pertinent et d'actualité et notre choix est tombé sur l'article :
« **Classification et conception spécifique de l'amorce pour la détection précise du SARS-CoV-2 à l'aide du deep learning ».** Publié le 13 janvier 2021 et réalisée par une équipe de chercheurs d'Utrecht une commune et ville néerlandaise, à la tête de cette équipe **Alejandro Lopez-Rincon** un chercheur de l'institut pharmaceutique et scientifique d'Utrecht.

Nous allons dans ce qui suit décortiquer cet article en évoquant les méthodes et techniques d'intelligence artificielle utilisées ainsi qu'aux résultats aboutis pour la découverte de séquences génomiques représentatives dans le SARS-CoV-2.

II. Problématique :

a. Introduction :

En décembre 2019, le SARS-CoV-2, un nouveau coronavirus infectant l'homme, a été identifié à Wuhan, en Chine, à l'aide du séquençage de prochaine génération (NGS). Au 12 août 2020, le nouveau SARS-CoV-2 compte 20 162 474 cas confirmés dans presque tous les pays. De plus, le SARS-CoV-2 a un taux de mortalité estimé entre 3 et 4 %, et il se propage plus rapidement que le SARS-CoV et le MERS-CoV.

La famille des Coronaviridae[1] présente un sens positif, génome[2] d'ARN à un brin. Ces virus ont été identifiés chez les hôtes aviaires et mammifères, y compris les humains.

[1] Les Coronaviridae sont une famille de virus à ARN simple brin enveloppés, de sens positif.
[2] Génomes est l'ensemble du matériel génétique d'une espèce codé dans son acide désoxyribonucléique (ADN), Il contient en particulier tous les gènes codant des protéines ou correspondant à des ARN structurés. Il se décompose donc en séquences codantes.

Dans le cas spécifique du SARS-CoV-2, des tests RT-qPCR utilisant des amorces[3] dans les gènes ORF1ab et N ont été utilisés pour identifier l'infection chez l'homme.

Mis à part les problèmes de test faussement négatifs, les tests SARS-CoV-2 peuvent donner une petite partie des faux positifs grâce à la détection non spécifique d'autres coronavirus, car le virus est étroitement lié à d'autres organismes coronavirus. En outre, le SARS-CoV-2 peut être présent avec d'autres infections respiratoires.

D'où la nécessité d'améliorer les outils de diagnostic existants pour contenir la propagation.

b. Techniques utilisées :

Des outils de diagnostic combinant des tomodensitométries et le deep learning ont été proposés, ce qui a amélioré la précision de détection de 82,9 %, l'une des méthodes utilisées est la classification à l'aide de techniques de séquençage viral, elle est principalement basée sur des méthodes d'alignement telles que **BLAST**[4]. Ces méthodes reposent sur l'hypothèse que les séquences DNA partagent des caractéristiques communes, et leur ordre prévaut entre les différentes séquences. Toutefois, ces méthodes souffrent de la nécessité d'avoir besoin de séquences de base pour la détection et parfois plusieurs tests sont nécessaires pour avoir un diagnostic précis.

Par conséquent, comme alternative, des méthodes du deep learning ont été suggérées pour la classification des séquences d'ADN. L'avantage de ces méthodes est qu'elles n'ont pas besoin de fonctionnalités présélectionnés pour identifier ou classer les séquences d'ADN.

La méthode consiste à l'utilisation d'un CNN (convolutional Neural Networks) pour séparer les coronavirus appartenant à différentes souches, y

[3] L'amorce est une courte séquence d'ARN ou d'ADN, complémentaire du début d'une matrice, servant de point de départ à la synthèse du brin complémentaire de cette dernière matrice par une ADN polymérase.
[4] BLAST st une méthode de recherche heuristique utilisée en bio-informatique. Il permet de trouver les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés, et de réaliser un alignement de ces régions homologues.

compris le SARS-CoV-2, nous appliquons des techniques inspirées par XAI dans la vision par ordinateur pour découvrir des séquences représentatives de l'ADNc que le réseau utilise pour classer le SARS-CoV-2.

Nous validons ensuite les séquences découvertes sur des ensembles de données non utilisés lors de la formation du CNN, et montrons comment les exploiter pour créer un nouvel ensemble très instructif de fonctionnalités de séquence (par exemple des séquences virales).

De telles séquences peuvent ensuite être inspectées et analysées par des experts humains. Les résultats expérimentaux montrent que le nouvel ensemble de caractéristiques de séquence conduit les classificateurs traditionnels et simples à évaluer correctement le SARS-CoV-2 avec une précision remarquable (> 99 %). Quelques-unes des séquences découvertes possèdent également les caractéristiques correctes pour devenir des amorces, car il suffit de vérifier leur présence dans les échantillons pour identifier spécifiquement le SARS-CoV2.

Les essais en laboratoire sur les séquences les plus prometteuses identifiées, ont montré que les amorces trouvées par notre approche peuvent être une alternative viable aux amorces couramment adoptées au moment de la rédaction. Ces résultats pourraient ouvrir la voie à une procédure automatique pour la conception des amorces, voir Fig. 1 pour le flux de travail proposé.

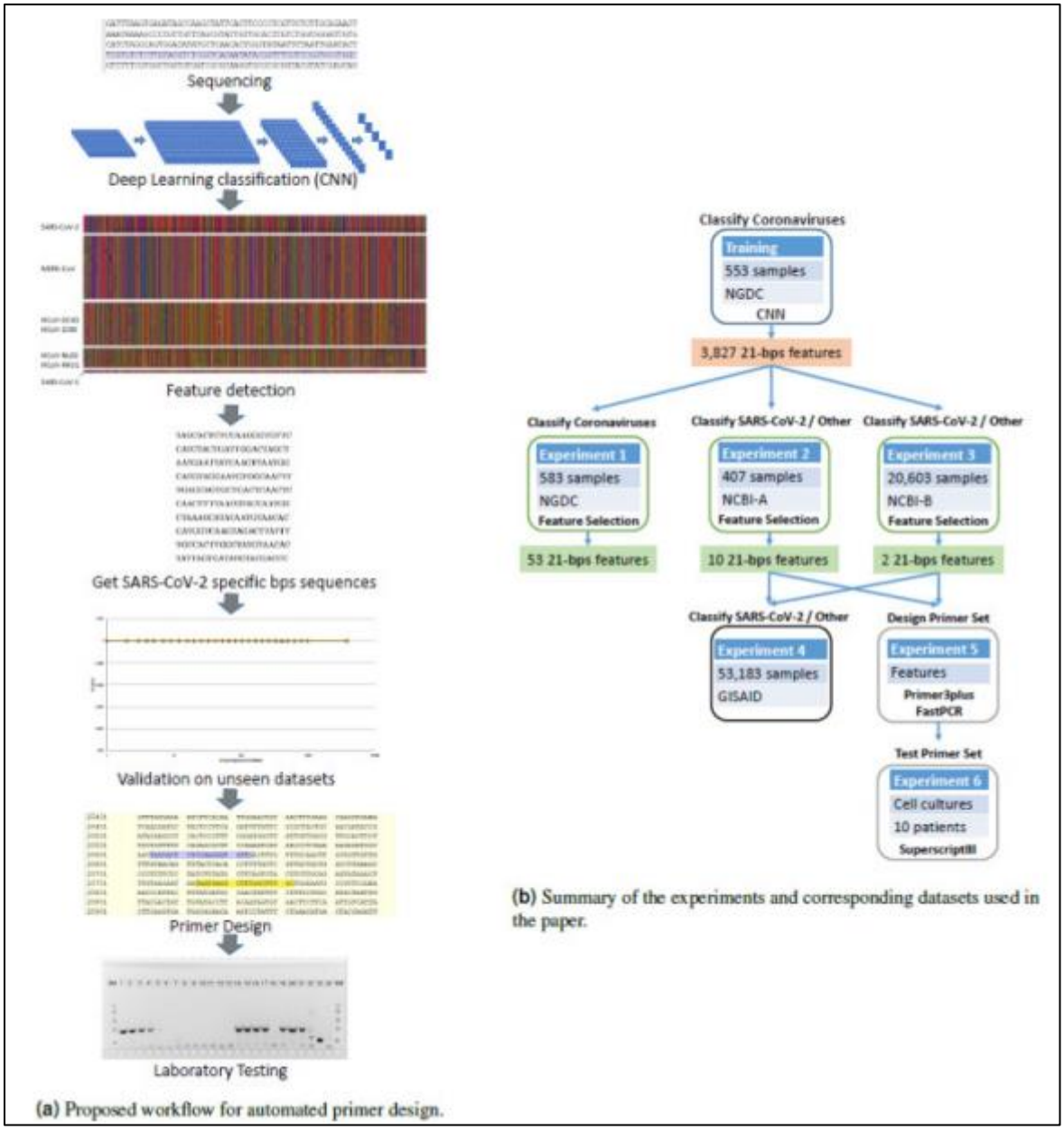

Figure 1 : (a) montre le flux de travail proposé pour la conception automatisée des amorces pour les virus. À droite, (b) résume les différentes expériences rapportées dans le document, ainsi que les ensembles de données utilisés dans chaque essai.

Le CNN utilisé pendant toutes les expériences est composé d'une couche convolutionnelle avec 12 filtres ou poids différents (chacun avec la taille de la fenêtre 21, et un rembourrage uniforme de 10 étapes de chaque côté) avec maxpooling (taille de la piscine 148 et foulée 1), une couche entièrement connectée (196 unités linéaires rectifiées avec probabilité d'abandon 0,5), et une couche softmax finale avec 5 unités, pour différencier les différentes classes de souches de coronavirus. L'optimiseur utilisé est Adaptive

Momentum (ADAM[5]), avec un taux d'apprentissage et une taille de lot de 50 échantillons, courir pendant 1000 époques. Un résumé graphique du CNN utilisé dans les expériences est rapporté dans Fig. 2.

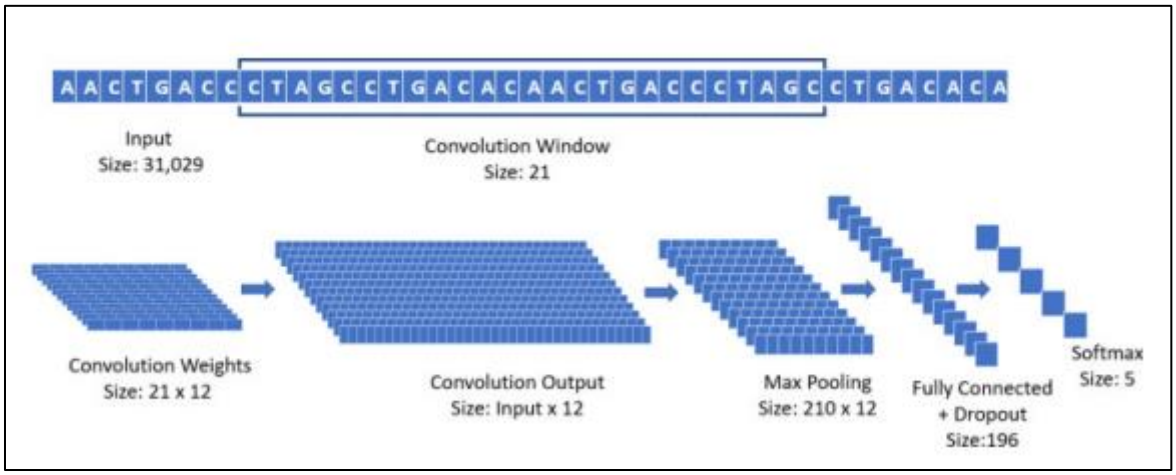

Figure 2 : Représentation graphique de l'architecture du CNN utilisée dans les expériences.

c. Résultats :

Le CNN formé décrit précédemment a obtenu une précision moyenne de 98,73% dans une validation croisée 10 fois stratifiée. Observation de la matrice de confusion pour les 5 classes considérées, il est remarquable de constater que même les échantillons provenant de classes sous-représentées étaient pour la plupart correctement positionnés. Un tel résultat encourageant peut indiquer que le réseau était régulièrement en mesure de découvrir des séquences significatives pour séparer les différentes classes de virus.

Récapituler les résultats des expériences, nous avons découvert 12 séquences significatives de 21 points de base qui caractérisent le mieux le SARS-CoV-2. Pour toutes les données analysées, ces séquences n'apparaissent que dans les échantillons du SARS-CoV-2 et non dans d'autres virus, Fait remarquable, nos résultats surpassent les publications antérieures en utilisant l'apprentissage automatique pour identifier le SARS-CoV-2.

La méthode 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), nos résultats démontrent

[5] Adam est une méthode de descente de gradient stochastique qui calcule les taux d'apprentissage adaptatif individuels pour différents paramètres à partir d'estimations des moments de premier et de second ordre des gradients.

clairement la puissance de notre méthode pour sélectionner des séquences potentielles pour une validation plus approfondie.

d. 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 taux 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 leurs limites.
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[6], NCBI[7] et le GISAID[8].
5. Le principal travail du CNN est de séparer les coronavirus appartenant à différentes souches, y compris le SARS-CoV-2. On génère ensuit des séquences représentatives de l'ADNc que le réseau utilise pour classer le SARS-CoV-2. Après validation des séquences découvertes les résultats montre que les classificateurs traditionnels et simples à évaluer correctement le SARS-CoV-2 avec une précision remarquable (> 99 %).

[6] Le référentiel du National Genomics Data Center (NGDC)
[7] Le référentiel du National Center for Biotechnology Information (NCBI)
[8] Le référentiel de l'Initiative mondiale sur le partage de toutes les données sur la grippe (GISAID)

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 Booléenne (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 SARS-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 SARS-CoV-2.
10. Exemple dans l'article d'une séquence qui se trouve qu'à l'intérieur de la classe du SARS-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.

III. Sources et références :

Ma seule et unique source pour ce TP1 est l'article « **Classification et conception spécifique de l'amorce pour la détection précise du SARS-CoV-2 à l'aide du deep learning ».** Publié le 13 janvier 2021 et réalisée par une équipe de chercheurs d'Utrecht, à la tête de cette équipe **Alejandro Lopez-Rincon** un chercheur de l'institut pharmaceutique et scientifique d'Utrecht. Lien de l'article : Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning - PubMed (nih.gov)

L'article lui-même s'inspire de plusieurs références :

- Woo PC, Huang Y, Lau SK, Yuen K-Y. Coronavirus genomics and bioinformatics analysis. Viruses. 2010;2:1804–1820. doi: 10.3390/v2081803. - DOI - PMC - PubMed
- Lu R, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. Lancet. 2020;395:565–574. - PMC - PubMed
- World Health Organization . WHO Report Coronavirus Disease 2019 (COVID-19) Geneva: World Health Organization; 2020.
- Wang Y, Kang H, Liu X, Tong Z. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. J. Med. Virol. 2020;20:20. - PMC - PubMed
- Corman VM, et al. Detection of 2019 novel coronavirus (2019-ncov) by real-time RT-PCR. Eurosurveillance. 2020;25:20. - PMC – PubMed.