Drug Design et Coronavirus: Intelligence artificielle et repositionnement du médicament.

Drs Chiraaz Maleej [a,*] Aurélien Jobard [a,*] Amor Hosni [a,*] et Sébastien Mourey [a,*]

DU Intelligence artificielle en Santé CHU Bourgogne mai 2021

La crise sanitaire provoquée par la pandémie de Coronavirus a accéléré le développement des méthodes dites de « drug repurposing » face aux méthodes de drug design traditionnelles. `Cette méthode consiste à identifier, pour une pathologie donnée, les cibles biologiques (serrures) afin de modéliser des ligands (clés) ayant déjà fait l'objet d'essais sur l'homme qui pourraient exercer une nouvelle action thérapeutique. Dans la course à la recherche d'un traitement anti-COVID il existe 2 options : la première consiste à créer in silico un ligand capable de se fixer sur les cibles impliquées dans le cycle viral et la deuxième consiste à screener des molécules existantes et de retenir le meilleur candidat via la meilleure affinité, c'est ce qu'on appelle le repositionnement (Drug repurposing). Pour ce projet, nous avons opté pour la deuxième option qui présente l'avantage d'être moins chère, plus rapide et offre la possibilité d'appliquer l'intelligence artificielle (IA) pour optimiser les résultats.

INTRODUCTION

La conception de médicaments se base essentiellement sur une complémentarité structurelle entre le ligand (la molécule) et la cible biologique. L'étape clé consiste à élucider les mécanismes impliqués dans une maladie pour identifier des cibles biologiques (récepteurs, enzymes ou autres protéines). Un pharmacophore (portion de la molécule responsable de l'activité pharmacologique) susceptible d'interagir avec les cibles sera défini. Il suit l'étape de screening sur l'ensemble de molécules, d'une base de données, possédant ce pharmacophore et une étape finale de tri en se basant sur un critère au choix (affinité, homologie...) pour ne conserver que les molécules disposant d'un véritable potentiel thérapeutique.

Le modèle actuel de R&D pour les produits pharmaceutiques est long, coûteux et affecté par une mauvaise exploitation des données disponibles se traduisant par un gaspillage de ressources. L'intérêt d'utiliser l'IA au profit du drug design permettrait d'accélérer l'identification des substances prometteuses qui pourraient devenir des candidats médicaments en optimisant le screening en termes de durée et de coût. Par ailleurs, l'utilisation des algorithmes permet d'augmenter la sélectivité tout en réduisant la toxicité des candidats.

MÉTHODOLOGIE

1) Qualification des bases de données

Le projet a débuté par une étape d'analyse des différentes bases de données (Uniprot, DrugBank, Protein DataBank, ZINC, PubChem…) mises à disposition en open source afin de se familiariser avec leur architecture, les manipuler et identifier la disponibilité des données qui vont nous aider dans la recherche de nos molécules candidats anti-COVID.

2) Identifier les cibles pharmacologiques liées au Covid

La deuxième étape nous a permis d'identifier, par recherche bibliographique, à travers une compréhension du cycle de vie du coronavirus, les différents mécanismes d'action potentiels. On décrit alors une action directe (sur la structure même du virus), et une action indirecte (sur les symptômes et la réaction inflammatoire induite par le virus).

Nous avons pu ainsi identifier différentes protéines structurelles (Spike, Membrane, Enveloppe et Nucléocapside) mais également des protéines et enzymes impliquées tout au long du processus de fixation, d'intégration et de réplication du virus. Par ailleurs, nous avons fait le choix de cibler la réaction inflammatoire de l'hôte suite à son infection par le virus en s'attaquant ainsi aux étapes clés du cycle viral lytique et à la réponse immunologique due au virus.

L'analyse bibliographique à ce sujet nous a permis d'identifier plusieurs classes thérapeutiques pouvant avoir une action pharmacologique anti-COVID. Les essais cliniques en cours utilisant ces classes thérapeutiques nous ont permis de valider l'intérêt de ces molécules pour cette indication.

3) Screener les molécules actives disposant d'une AMM

Cette étape à été réalisée, en parallèle, par deux équipes (Bio-team et IT-team) dans le cadre du Datathon. Le protocole appliqué peut être décrit comme suit:

- Extraction de la base de données DrugBank -> 70K fichiers identifiés sous format XML (Utilisation MV Amazon).

- Renommage du fichier XML par son identifiant DrugBank ID.
- Pour chaque ID DrugBank, extraction des drug interactions PUIS extraction de chaque molécule associée à ces interactions.
- Génération d'un fichier csv comprenant notamment les molécules ayant été approuvées (AMM).
- Création d'une interface graphique sous Python avec algorithmie simple orientée objet.

Cette extraction des données nous a permis d'identifier automatiquement les ligands permettant d'interagir avec les principales cibles actives que nous avons sélectionnées et générer une première chimiothèque de molécules. Il s'agissait alors de ne sélectionner que les candidats ayant la meilleure affinité pour la cible et d'écarter celles dont la toxicité est avérée. A la fin de cet exercice, les résultats des deux équipes ont été comparés et les candidats intéressants ont été analysés (vérification de leur statut, objet d'essais en cours...etc). Un tri nous a permis de ne garder que les molécules ayant déjà fait leurs preuves et ayant obtenu leur Autorisation de Mise sur le Marché. Au final, une short-list avec les molécules ayant validées tous nos critères de sélection a été établie et le résultat ainsi interprété pour évaluer la précision et l'efficacité de notre algorithme.

RÉSULTATS

Analyse des cibles directes (Cycle viral lytique)

Le coronavirus possède à sa surface des protéines dites « spikes » qui nécessitent d'être activées par la furine, une enzyme présente dans notre organisme (priming) pour ensuite s'accrocher sur le récepteur ACE2 (présent sur de nombreuses cellules de l'organisme dont le poumon et impliqué, notamment, dans la régulation de la tension artérielle) et pouvoir pénétrer dans les cellules hôtes par endocytose. Après la fusion et la pénétration, aura lieu le largage de l'ARN viral dans le cytoplasme, et la lecture traductionnelle par les ribosomes. La lecture d'un seul de cet ARN va synthétiser à la fois les protéines structurelles du virus ainsi des enzymes permettant de dupliquer l'ARN et conduire à la formation de nouveaux virions. Finalement les brins d'ARN synthétisés sont combinés avec la protéine N formant la nucléocapside et l'assemblage avec les glycoprotéines d'enveloppe permet le bourgeonnement de nouvelles particules virales (1).
La première stratégie thérapeutique directe vise à empêcher le virus de pénétrer dans la cellule en jouant sur les mécanismes nécessaires à la fixation du virus à son récepteur, son endocytose ou la fusion membranaire. Un autre mode d'introduction se ferait via le corécepteur TMPRSS2 très présent sur les pneumocytes et qui introduit le virus par fusion de sa membrane lipidique peut faire l'objet d'une action thérapeutique anti-COVID.

Priming	Fixation	TMPRSS2
Héparines (Inhibiteur Furine)	Chloroquines et dérivés (anti Rp ACE)	camostat mesylate

Cibler les protéases virales indispensables à la réplication avec les antis VIH (Lopinavir, Ritonavir) consiste une piste thérapeutique qui a été largement documentée par les publications scientifiques.
D'autres hypothèses telles que l'inhibition par IPP à forte dose (oméprazole et dérivés) induisant un déséquilibre du PH intra-vacuolaire semble limiter l'infection.

Analyse des cibles indirectes (Réponse immunologique)

L'infection des cellules épithéliales et immunitaires du tractus respiratoire génère plusieurs signaux qui vont ensuite activer des facteurs de transcription et entraîner la sécrétion de cytokines (TNF-α, IL-1, IL-6) et l'attraction de cellules inflammatoires, et d'interférons de type I (IFN-1). Ce phénomène hyper-inflammatoire, conséquence d'une réponse immunitaire disproportionnée est nommé «tempête hyper-inflammatoire" ou "Orage des Cytokines". Cette voie est centrale dans la réponse antivirale initiale, et permet notamment d'inhiber la réplication virale, de protéger les cellules non-infectées et de stimuler l'immunité lymphocytaire antivirale (lymphocytes T CD8, NK) conduisant à la lyse des cellules infectées.

Des cibles dirigées contre le récepteur de l'interleukine-6 (IL-6), IL-1 et des TNF apparaissent comme un moyen prometteur d'arrêter la tempête cytokinique.

IL1	IL6	TNF Alpha
Canakinumab (Ilaris)	Tocilizumab Siltuximab Sarilumab	Infliximab Adalimumab Golimumab Centolizumab Etanercept Thalidomide Hydroxythalidomide Pomalidomide Acide glycyrrhizique

Par ailleurs, l'inhibition de cette cascade en amont par la voie des MAP kinases p38 aura une justification solide comme stratégie thérapeutique.

Cibles Identifiées par IA

Le travail réalisé par les deux équipes au cours du Datathon nous a permis d'établir cette liste de molécules candidates.

IL1	IL6	TNF Alpha	Tubuline
VX-702 **Binimetinib** Etiprednone etylacetate Cefazoline	Dilmapimod **Binimetinib** **VX-702**	Pranlukast Thalidomide Dilmapimod **Binimetinib** **VX-702**	Albendazole Colchicine Estramustine Tetracycline Lansoprazole

A notre connaissance, les derniers essais et publications portant sur le VX-702 (molécule expérimentale) remontent à 2005 sans suite favorable

Binimetinib un inhibiteur de MEK (Mitogen activated protein Kinase) est indiqué en cas de mélanomes métastatiques avec des mutations spécifiques. Il fait actuellement l'objet d'un essai clinique phase II en association avec encorafenib chez des patients atteints d'un cancer à cellules non caractéristiques (petites tailles) présentant une mutation BRAF V600E.

DISCUSSION

En évaluant le résultat livré par IA, nous retrouvons une grande complémentarité avec ce que nous avons trouvé par recherche bibliographique à cible identique. Cependant, nous avons eu des résultats aberrants type Ginseng ou encore des molécules manquantes et attendues comme la Colchicine. Plusieurs points d'amélioration sont discutés ci-dessous :

1. *Non utilisation des réseaux neuronaux.*

En effet, en vue du faible nombre de molécules ressortant du pré-screening, soit une cinquantaine, le deep learning n'aurait pas été efficace avec un tel échantillon. En revanche, pour de plus grands nombres de molécules pré-screenées, en utilisant un set de données d'apprentissage pour un output de type Récompense, l'utilisation de l'IA aurait eu plus de sens à ce niveau. Ainsi, les machines virtuelles d'Amazon n'ont pas été utilisées à leur plein potentiel mais auront servi toutefois à paralléliser le traitement des données et donc de gagner du temps.

2. *Problème de qualité de la base de données.*

La base de données DrugBank a montré un certain nombre de limites notamment sur les relations bijectionnelles entre des molécules lorsqu'on parle d'interactions. Par ailleurs, l'algorithme utilisé au cours du challenge "Datathon" utilise des boucles If relativement simple et ne prend pas en compte la sémantique de type "Inhibition" ou "Stimulation". Une piste d'amélioration serait alors de créer initialement un dictionnaire de mots définis par l'utilisateur portant sur la sémantique de ce qu'on souhaite trouver. Les résultats pourraient ainsi ressortir avec une meilleure précision. Enfin, les algorithmes et les filtres utilisés sont sensibles à la casse dont notamment les accents. Ce problème peut ainsi réduire la qualité des résultats.

3. *Hypothèse sur des "outliers" à travers l'exemple du Ginseng.*

Une brève revue de littérature nous informe que le Ginseng est un phytothérapeutique qui pourrait avoir une influence dans les maladies auto-immunes notamment en modulant les différents niveaux de cytokine et donc de manière générale la réponse inflammatoire (https://doi.org/10.1016/j.jgr.2018.10.002). De même, notre algorithme se base principalement sur la "Drug Interaction" et les cibles initiales choisies initialement sont des cytokines impliquées dans la réponse pro-inflammatoire. Ainsi, selon la qualité des données, l'interaction entre le Ginseng et des cytokines a peut-être été établie ce qui a ainsi fait ressortir cette molécule dans le pré-screening.

CONCLUSION

Ce premier travail nous a offert l'opportunité d'aborder l'IA appliquée au Drug design sur le plan pratique. Une première expérience enrichissante qui nous a donné une vue d'ensemble sur la conduite d'un projet de développement d'un algorithme en combinant les compétences scientifiques et informatiques.

Dans la continuité, nous poursuivrons notre aventure avec un nouvel objectif qui consiste à construire un modèle d'IA permettant d'identifier les molécules réduisant ou aggravant la durée d'hospitalisation en vie réelle (1 an de recul).

1. Récupérer la base de données patients covid-19 avec leurs données des traitements concomitants et la durée d'hospitalisation. (Un dépôt de dossier est en cours auprès du SNDS)
2. Classifier selon la méthode naïve Bayésienne sur la durée d'hospitalisation (ou décès) en fonction de la classe d'âge et des co-morbidité associées en 4 sous-groupes : (<,Moy,> et dcd)
3. L'objectif sera alors de faire ressortir les molécules (patterns) les plus représentées statistiquement via un algorithme non supervisé. Par exemple, supposons une molécule X présente 10 fois en tout mais apparaissant 8 fois dans le groupe favorable, X ressortira comme ayant un impact favorable sur la Covid.
4. En comparant nos données avec l'actualité des essais cliniques nous pourrons présupposer la viabilité d'un essai en cours ou orienter de nouvelles pistes.