PROPOSITION DE SUJET DE THESE

SIGLE ET NOM DU LABORATOIRE : BIOLOGIE FONCTIONNELLE ET ADAPTATIVE, CNRS 8251-INSERM ERL U1133
NOM DE L'EQUIPE : COMPUTATIONAL PHARMACOLOGICAL PROFILING
DIRECTEUR DE THESE : CAMPROUX ANNE-CLAUDE
ADRESSE : CASE 7113, 35 RUE HELENE BRION, UNIVERSITE PARIS CITE, 75205 PARIS CEDEX 13

TITRE DE LA THESE : AMELIORATION DES METHODES IN SILICO DE DRUG DESIGN (CONCEPTION DE MEDICAMENTS ASSISTEE PAR ORDINATEUR) PAR LA PRISE EN COMPTE DES STRUCTURES PROTEIQUES ET DE LEURS FLEXIBILITES

PRESENTATION DU SUJET

1. *Le contexte scientifique du projet*

Dans les pipelines de découverte de médicaments, il est bien connu qu'un médicament peut-être impliqué dans différentes fonctions pathologiques et interagir avec plusieurs cibles. Cette observation relève du concept de la polypharmacologie [1], c'est-à-dire la capacité d'un médicament à se lier à plusieurs cibles. Celle-ci peut avoir des effets bénéfiques, tels que la reconsidération des médicaments, ou des effets indésirables hors cible [2] entraînant des réactions médicamenteuses adverses ou des effets secondaires [3].

La conception de médicaments assistée par ordinateur (Computer Aided Drug Design, CADD) est une discipline permettant de rationaliser et de réduire le coût du temps de recherche de candidats médicaments pour un large éventail de maladies [4]. Avec les améliorations des techniques modernes de détermination de la structure, les informations structurelles des protéines cibles sont de plus en plus disponibles et utilisées pour prédire les interactions protéine-médicament [5] et les méthodes de CADD basées sur la structure tridimensionnelle (3D) des cibles thérapeutiques [6] (Structure-based Drug Design, SBDD) sont en pleine expansion [7]. Les interactions entre les protéines et les médicaments se produisent au niveau des sites de liaisons des protéines, qui correspondent à de petites cavités, appelées poches de liaison du ligand. Celles-ci répondent à des propriétés géométriques et physico-chimiques particulières, permettant ainsi des interactions spécifiques. La caractérisation fine de ces poches de liaison capables de fixer des molécules médicaments et de leur flexibilité doit permettre l'amélioration des méthodes SBDD telles que les approches de criblage virtuel et l'amarrage moléculaire (docking).

Le criblage virtuel est une technique de calcul utilisée avec succès pour identifier des ligands pour des protéines cibles à partir de grandes bases de données de molécules virtuelles [8]. L'amarrage moléculaire est un processus de calcul largement utilisé pour prédire rapidement les modes de liaison et les affinités de ligands avec leurs protéines cibles [9]. La prédiction de la position, orientation et conformation d'un ligand lorsqu'il est lié à une protéine cible est appelée « pose » de liaison du ligand. Les techniques d'amarrage moléculaire sont associées à deux types de fonctions : des fonctions de score de docking (ou de classement) pour classer des centaines de poses de liaison possibles de ligand afin de sélectionner les plus favorables, et également des fonctions de score d'affinité pour prédire l'affinité de liaison protéine-ligand. Ces fonctions de score ont été largement développées au cours des dernières décennies et celles récemment développées à partie d'approches de Machine Learning (ML) surpassent les fonctions classiques disponibles dans les logiciels d'amarrages [10-12].

Malgré des applications réussies de CADD à la conception de médicaments moderne, ces méthodes présentent différentes limites [13,14].

(i) L'augmentation du nombre de données engendre un coût computationnel important des approches de criblage virtuel qui testent tous les ligands des bases de données sur des protéines cibles, alors qu'un très petit sous-ensemble de composés les mieux classés sera pris en compte pour une évaluation expérimentale et la plupart des molécules criblées sont rejetées[15].

(ii) Un défi dans l'amarrage protéine-ligand est la prise en compte de la flexibilité moléculaire des partenaires, car la protéine et le ligand changent de conformation lors de la liaison. En ce qui concerne l'amarrage flexible des ligands, les programmes d'amarrage permettent d'obtenir de meilleures poses prédites, mais ont des difficultés à classer ces poses parmi les meilleures. Ces résultats suggèrent que les défis posés par les algorithmes actuels d'amarrage de ligands flexibles concernent la fonction de score de docking c'est-à-dire la méthode de classement des poses [16]. La prise en compte de flexibilité des protéines est un problème beaucoup plus difficile à résoudre que la flexibilité des ligands lors de l'amarrage en raison du nombre beaucoup plus élevé de degrés de liberté dans les protéines [16]. Les techniques d'amarrage protéine-ligand les plus largement utilisées supposent une protéine rigide (c'est-à-dire que les positions de tous les atomes de protéine sont fixes). L'amarrage avec une protéine rigide échoue souvent à produire une pose de ligand lorsque la forme de la poche de liaison de la protéine doit s'adapter au ligand à lier, cela a pour conséquence de classer ces poses défavorablement. Une variété de techniques d'amarrage flexible des protéines tente de résoudre ce problème en permettant à un nombre limité de résidus de la poche de liaison de la protéine de se déformer pendant l'amarrage[17], cependant cette approche est très coûteuse en calculs et en temps[18].

2. *Les questions posées*

Ce travail de thèse a pour but d'intégrer les caractérisations des sites de liaison et de leurs flexibilités développées par l'équipe afin de contrer les limites des approches de criblage virtuel et l'amarrage moléculaire (docking), et d'améliorer les approches de SBDD et la prédiction des ligands partenaires d'une protéine cible d'intérêt.

(i) Dans le but de réduire le coût computationnel du criblage virtuel, le premier objectif sera de développer une méthode permettant la recherche de poches similaires de notre protéine cible d'intérêt à partir d'une base de données en intégrant la flexibilité des poches. Ces poches similaires permettront de concevoir une bibliothèque de ligands restreinte et privilégiée à notre protéine d'intérêt.

(ii) Une fois la bibliothèque de ligands construite, le deuxième objectif sera de réaliser de l'amarrage flexible avec ces ligands. De plus, afin d'éviter de réaliser l'amarrage flexible sur la protéine entière, nous allons le réaliser en nous limitant aux résidus clés de la poche d'intérêt de la protéine, déterminés grâce aux connaissances des poches du laboratoire [19], ce qui va permettre de prendre en compte la flexibilité de la protéine avec un coût en calculs beaucoup moins important.

(iii) Le troisième objectif sera d'améliorer les fonctions de score de ML récentes en ajoutant des paramètres pour prendre en compte la flexibilité afin de mieux classer les poses d'amarrage.

Une fois ces différentes méthodes développées, elles pourront être appliquées aux travaux récents de notre équipe [19]. Ceux-ci ont consisté à rechercher des cavités protéiques susceptibles d'héberger un candidat-médicament capable d'inhiber l'interaction entre le domaine RBD de la protéine NS1 du virus Influenza A et l'ARN sur ces différentes structures et sous-types. Notre dernier objectif sera donc de déterminer de bons candidats médicaments pour ces poches identifiées.

3. *Les sources de données qui seront utilisées*

Grâce aux technologies récentes permettant de déterminer les structures d'un grand nombre de protéines telles que la microscopie cryoélectronique (EM), la résonance magnétique nucléaire (RMN), la cristallographie aux

rayons X, et des méthodes informatiques (modélisation par homologie, AlphaFold…), un nombre croissant de structures protéiques tridimensionnelles sont disponibles (plus de 204.000) dans la « Protein Data Bank » (PDB) [20] et de grandes bases de données de haute qualité sur les complexes protéine-médicament telles que MOAD [21], PDBBind [22], BindingDB [23], mais également des bases de données de ligands telles que ZINC [24] et SMDC [25].

De grands ensembles de données de plus de 225.000 poches en complexes et leurs médicaments correspondants ont été extraits des complexes disponibles dans la PDB au cours du doctorat de Mme Cerisier [26].

Des outils de bio-informatique structurelle (logiciels d'amarrage : AutoDock [27], Glide [28], …) et divers descripteurs tels que les pharmacophores [29] seront utilisés pour caractériser les poches et leur flexibilité.

4. *Les méthodes*

Des méthodes d'apprentissage automatique non supervisé telles que la classification hiérarchique, la classification Kmeans, Analyse en Composantes Principales (PCA) …

Diverses méthodes d'apprentissage automatique supervisé telles que CART, Random Forest, l'intelligence artificielle (Machine Learning) et les approches d'apprentissage profond (Deep Learning) seront utilisées. Les descripteurs pour déterminer les similarités des poches seront sélectionnés par différentes méthodes d'apprentissage automatique afin de prendre en compte la flexibilité des poches.

Diverses méthodes de fonctions de score seront étudiées : les fonctions dites classiques [27,28], et celles basées sur le Machine Learning et sur le Deep Learning[11,12].

5. *Le calendrier prévisionnel*

- Octobre 2023 - Mars 2024 : Bibliographie, appropriation des méthodes de bio-informatique structurales (amarrage, ...) ; Exploration des similarités des poches protéiques en tenant compte de la flexibilité
- Avril 2023 - Septembre 2024 : Bibliographie des fonctions de score existantes ; Docking flexible ; Création d'une fonction de docking en ajoutant des paramètres de flexibilité.
- Octobre 2024 - Mars 2025 : Création d'une fonction de notation en ajoutant des paramètres de flexibilité ; Tests des nouvelles fonctions de score sur des jeux de données.
- Avril 2025 - Septembre 2025 : Application des différentes méthodes élaborées sur NS1
- Octobre 2026 - Mars 2026 : Extension du protocole à de nouvelles cibles
- Avril 2026 - Septembre 2026 : Élaboration du manuscrit de thèse, soumission à l'examinateur et soutenance

6. *Le thème de chacun des articles prévus*

Publication 1 : Protocole de filtrage statistique pour la sélection d'une banque de ligands candidats pour une protéine cible

Publication 2 : Nouvelle fonction de docking permettant l'amélioration de l'amarrage flexible du ligand : application à la recherche de candidats médicaments capables d'inhiber l'interaction entre le domaine RBD de la protéine NS1 du virus Influenza A et l'ARN étudiée par l'équipe [19].

7. *Bibliographie*

1. Lavecchia et al. In silico methods to address polypharmacology: current status, applications and future perspectives. Drug Discov Today 21(2): 288–298 (2016)
2. Zhou et al. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. Scientific reports, 1-13 (2015)
3. Abi Hussein et al. System Biology: a new paradigm for drug discovery. The Practice of Medicinal Chemistry, 409–425 (2015)
4. Baig MH, Ahmad K, Roy S, et al. Computer Aided Drug Design: Success and Limitations. Curr Pharm Des. 2016;22(5):572-581.
5. Mestres et al. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. Molecular bioSystems, 5(9):1051–1057 (2009)

6. Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Sci. 1998, 7, 1884–1897.
7. Lounnas V., Ritschel T., Kelder J., McGuire R., Bywater R.P., Foloppe N. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. Comput. Struct. Biotechnol. J. 2013;5:e201302011.
8. Lionta E., Spyrou G., Vassilatis D.K., Cournia Z. Structure-based virtual screening for drug discovery: Principles, applications and recent advances. Curr. Top. Med. Chem. 2014;14:1923–1938.
9. Waszkowycz B., Perkins T.D.J., Sykes R.A., Li J. Large-scale virtual screening for discovering leads in the postgenomic era. IBM Syst. J. 2001;40(2):360.
10. Su M. et al. (2019) Comparative assessment of scoring functions: the CASF-2016 update. J. Chem. Inf. Model., 59, 895–913.
11. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinformatics 26(9):1169–1175.
12. Wójcikowski, Maciej et al. "Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions." Bioinformatics (Oxford, England) vol. 35,8 (2019): 1334-1341.
13. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug Discov. Today. 2006;11(13-14):580–594.
14. Korb, Oliver et al. "Potential and limitations of ensemble docking." Journal of chemical information and modeling vol. 52,5 (2012):1262-74.
15. Shoichet, B. K. Virtual screening of chemical libraries. Nature 432, 862–865 (2004).
16. Sheng-You Huang, Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges, Briefings in Bioinformatics, Volume 19, Issue 5, September 2018, Pages 982–994
17. Miller EB et al. "Reliable and Accurate Solution to the Induced Fit Docking Problem for Protein-Ligand Binding." J Chem Theory Comput. 2021;17(4):2630-2639
18. Bender, Brian J et al. "A practical guide to large-scale docking." Nature protocols vol. 16,10 (2021): 4799-4832.
19. Sarah Naceri, Daniel Marc, Rachel Blot, Delphine Flatters, Anne-Claude Camproux. Druggable Pockets at the RNA Interface Region of Influenza A Virus NS1 Protein Are Conserved across Sequence Variants from Distinct Subtypes. Biomolecules, 2023, 13 (1), pp.64.
20. Berman et al. The protein data bank, Acta Crystallogr. Sect. D Biol. Crystallogr 58, 6, 899–907 (2002)
21. Ahmed et al. Recent improvements to Binding MOAD. Nucleic Acids Res. 43 (2015)
22. Liu, Zhihai; Su, Minyi; Han, Li; Liu, Jie; Yang, Qifan; Li, Yan; Wang, Renxiao *, "Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions", Accounts of Chemical Research, 2017, 50 (2): pp. 302-309.
23. Gilson, Michael K et al. "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology." Nucleic acids research vol. 44,D1 (2016): D1045-53.
24. Irwin JJ, Tang KG, Young J, et al. ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J Chem Inf Model. 2020;60(12):6065-6073.
25. Arkin MR, Ang KK, Chen S, et al. UCSF Small Molecule Discovery Center: innovation, collaboration and chemical biology in the Bay Area. Comb Chem High Throughput Screen. 2014;17(4):333-342.
26. PhD N. Cerisier, Modélisation des interactions protéines-effecteurs et développement de méthodes de prédiction des partenaires de ces interactions, november 2019.
27. Morris GM, Goodsell DS, Halliday RS, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem 1998;19:1639–62.
28. Friesner RA, Banks JL, Murphy RB, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47:1739–49.
29. Weill et al. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. J. Chem. Inf. Model. 50, 123-35 (2010)

PRÉREQUIS, FORMATION : BIOINFORMATIQUE STRUCTURALE, MÉTHODES DE DOCKING - SCORING, MACHINE LEARNING

CONTACT POUR CE SUJET : CAMPROUX ANNE
EMAIL : ANNE-CLAUDE .CAMPROUX@U-PARIS.FR
TELEPHONE : 0157278377

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Directeur : Pierre-Yves Boëlle
Responsable pour l'Université Paris Cité : Isabelle Boutron

SPECIALITE DE LA THESE

Santé publique - Epidémiologie	☐
Santé publique - Epidémiologie clinique	☐
Santé publique - Epidémiologie sociale	☐
Santé publique - Epidémiologie génétique	☐
Santé publique - Biostatistique	☐
Santé publique - Biomathématiques	☐
Santé publique - Biostatistique et Biomathématiques	☐
Santé publique - Informatique médicale	☐
Santé publique - Imagerie biomédicale	☐
Santé publique - Bioinformatique	☒
Santé publique - Recherches sur les services de santé	☐
Santé publique - Economie de la santé	☐
Santé publique - Science des données	☐
Santé publique – Prévention et promotion de la santé	☐

SIGNATURE DU. DE LA DIRECTEUR.TRICE DE THESE

[redacted]

VISA DU.DE LA DIRECTEUR.TRICE DU LABORATOIRE
(DEROGATION DE SIGNATURE NON ACCEPTEE)

AVIS FAVORABLE ☒

SIGNATURE

[redacted]

Ecole Doctorale 393
Centre Biomédical des Cordeliers
15, rue de l'Ecole de Médecine 75006 Paris
https://ed393.sorbonne-universite.fr/

Contact : ed393@sorbonne-universite.fr/Téléphone : 01.44.27.24.35