Equipe DSIMB, UMR-1134, Paris, (www.dsimb.inserm.fr)
Contact: catherine.etchebest@inserm.fr, julien.diharce@inserm.fr

Mots clés : Biomédicaments, bioinformatique, simulations, protéine, modélisation, prédiction, intelligence artificielle, apprentissage automatique, optimisation, analyse de données, programmation, informatique, biologie structurale, visualisation moléculaire, anticorps, agrégation.

Profil et compétences recherchés: Le candidat devra de préférence avoir obtenu un Master en bio-informatique ou sciences des données en biologie, ou biochimie théorique/modélisation moléculaire.

Pré-requis:
Connaissances avancées d'un langage de programmation
Connaissances de méthodes d'apprentissage & biostatistiques
Connaissances des concepts de la physico-chimie des biomolécules.
Connaissances des méthodes de simulations des biomolécules.

Qualités requises : Dynamisme, la curiosité, l'autonomie, l'imagination et de solides capacités d'adaptation. Le candidat devra en effet évoluer dans un environnement académique et industriel et répondre efficacement à ces différentes contraintes. Il devra faire preuve d'esprit d'équipe et être capable de communiquer par écrit et à l'oral en français et en anglais.

Résumé du projet de thèse

Au cours de la production d'un biomédicament, les industriels constatent des pertes significatives de produit actif. Ces pertes d'activité résultent des différentes étapes mises en œuvre, comme la purification ou le stockage. Les conditions environnementales, telles des variations de température, de pH jouent un rôle clé dans le maintien d'un produit actif.
La société Audensiel Healthcare s'est proposé de mettre en place un outil d'aide à la décision nommé AIDRUG pour la conception de biomédicaments, incluant le design de celui-ci jusqu'à la mise en place des conditions optimales de sa production industrielle. Le projet de thèse repose sur l'hypothèse que les pertes d'activité du produit sont dues à des changements structuraux et/ou dynamiques pouvant conduire à de la dénaturation ou à des phénomènes d'agrégation.
Il s'agira donc d'étudier finement les risques associés aux très basses températures rencontrées lors de la cryogénation du produit, à examiner l'impact de certains excipients ainsi que celui du pH. Une première étape consistera à *produire des données à l'aide des méthodes avancées d'échantillonnage conformationnel* et des méthodes adaptées pour *examiner l'effet du pH*. L'impact des différents facteurs sur les zones fonctionnelles sera exploré. *L'élucidation des mécanismes* responsables de changements dans ces régions permettra *d'identifier des*

descripteurs pertinents qui pourront être exploités lors de la seconde étape, qui consistera à *entraîner des méthodes d'apprentissage* afin de prédire l'effet observé et donc en retour à *proposer les paramètres optimaux limitant ce risque*.

Objectif et contexte

La mise en œuvre de l'outil AIDRUG s'appuie sur l'exploitation de données via des algorithmes d'apprentissage. Peu de données expérimentales sont cependant disponibles compte tenu de leur caractère confidentiel au niveau industriel. Afin de pallier cette carence, il est possible de simuler des données grâce à des outils performants de modélisation moléculaire. Lors d'une précédente thèse, des données ont été générées pour un anticorps thérapeutique, le pembrolizumab. L'intérêt s'est porté sur l'impact de la température sur la conformation de l'anticorps et le risque d'agrégation. Grâce à des simulations de dynamique moléculaire et la mise en œuvre d'outils d'analyse dédiés, il a été possible i) de montrer la très grande flexibilité de l'anticorps considéré et ii) d'identifier des états accessibles à certaines températures présentant un risque accru d'agrégation et donc un risque de perte d'activité. L'originalité de la démarche a consisté à estimer le risque grâce à une fonction combinant des propriétés structurales et dynamiques obtenues par simulation. Il a été ainsi possible de montrer que les outils actuels disponibles sont limités car ils ne considèrent que des changements conformationnels locaux ou tout du moins très restreints. Ils doivent donc être revisités afin de tenir compte des changements majeurs observés par les simulations.

S'appuyant sur ces résultats prometteurs mais encore préliminaires, la société Audensiel Healthcare, en partenariat avec l'équipe DSIMB de l'UMR-S1134, propose de poursuivre le développement de cette approche en affinant en particulier la fonction de score de risque et en élargissant l'étude à d'autres facteurs environnementaux influant l'activité d'un biomédicament.
Le projet de thèse comporte donc deux volets majeurs : i) un volet de recherche fondamentale visant à mieux comprendre les mécanismes moléculaires résultant en la perte de fonction du biomédicament dans différentes conditions ii) un volet à vocation industrielle consistant en la mise en œuvre de l'outil de prédiction. L'intrication de ces deux axes s'avère cruciale car c'est à travers l'analyse fine des résultats du volet i) que le volet ii) pourra être optimisé en sélectionnant des descripteurs les plus adaptés.
Les résultats obtenus lors de la thèse précédente ont montré que des risques importants de perte de fonction étaient liés aux basses températures, telles que celles rencontrées lors de la cryogénation du produit pour conservation. Lors de la première étape, il s'agira d'aller plus loin dans l'étude de ces basses températures, en examinant en particulier le rôle du solvant, pour finalement déterminer le meilleur cryoprotectant. La protéine d'intérêt sera à nouveau le pembrolizumab. L'impact sera évalué en premier lieu sur le maintien de la structure tertiaire (ou quaternaire si besoin) de la protéine en utilisant des méthodes avancées

d'échantillonnage conformationnel. Les contacts natifs, la stabilité de la structure tertiaire et secondaire seront précisément évalués car ce sont des prérequis pour maintenir la fonction. Les changements potentiels des propriétés de surface seront plus précisément considérés en particulier au niveau des sites de liaison. Seront ainsi examinées aussi bien les propriétés d'interaction du biomédicament avec sa cible que sa capacité d'autoassemblage irréversible conduisant à l'agrégation. La quantification de cette interaction sera réalisée à l'aide de calculs d'énergie libre. Des potentiels statistiques ou des approches basées sur des calculs de mécanique moléculaire seront utilisés. A ce jour, aucune étude de ce type n'a été menée.
Dans une seconde phase, sera examiné l'effet du pH. En effet, la gamme de pH visitée lors de l'étape de production d'un biomédicament s'échelonne entre 3 et 7 à 8 unités de pH. Une telle variation a une conséquence majeure sur les résidus titrables. La thèse précédente a montré que les risques d'agrégation impliquaient justement de tels résidus. Il s'agira donc d'évaluer pour quel pH les risques sont maximaux et à terme, proposer des modifications des résidus les plus sensibles.
Dans une troisième phase, l'impact d'agents chimiques par exemple des métaux ou d'acides gras, ajoutés lors des différentes phases de production sur la fonction sera lui-aussi étudié. La démarche sera là même que pour les deux phases précédentes.
En parallèle, une méthode basée sur des approches d'apprentissage basées sur l'intelligence artificielle sera mise en œuvre pour i) prédire les états conformationnels à partir de la structure 3D et ii) prédire les capacités d'autoassemblage. Le premier axe pourra être réalisé à l'aide d'une approche d'autoencodeur comme cela a déjà été réalisé pour d'autres systèmes protéiques. Le défi réside ici dans la taille du système et sa très grande flexibilité. Pour le second axe, l'apprentissage se fera sur les données d'autoassemblage produites par les simulations et visera à prédire le risque d'autoassemblage grâce aux données caractérisant l'implication de chaque résidu dans une interface protéine-protéine. La difficulté majeure réside dans l'identification des facteurs pertinents impliqués dans ce risque. La surface accessible au solvant est une mesure fréquemment exploitée ainsi que la nature du résidu considéré mais ces deux propriétés sont souvent insuffisantes pour définir finement le risque d'agrégation. Grâce à la nature des données produites, il sera possible d'élargir la nature des facteurs considérés et ainsi d'affiner la prédiction.
Cette procédure sera réalisée pour chacun des facteurs examinés. Enfin, une approche combinant ces différents facteurs sera développée.

Précision sur le financement et l'encadrement :

La thèse sera financée par un contrat CIFRE et se déroulera sur deux sites, celui du laboratoire académique (Paris) et celui de la Société Audensiel Healthcare (Proche banlieue). Le/la doctorant.e. sera affilié.e à l'école doctorale BioSPC de l'Université Paris Cité et sera encadré dans le laboratoire par la Prof. Catherine Etchebest et le Dr Julien Diharce

Conditions scientifiques matérielles et financières du projet de recherches

Pour la réalisation du projet de recherche, l'étudiant.e. bénéficiera d'un poste d'ordinateur fixe et portable lui permettant d'accéder à ses données à distance. Il pourra accéder à l'espace de stockage informatique réseau et au cluster de calcul de l'équipe et des centres nationaux. Des demandes d'heures de calculs seront effectuées chaque année. A noter que l'équipe a obtenu jusqu'à présent 6 Millions d'heures pour réaliser la première thèse.

Objectifs de valorisation des travaux de recherche du doctorant :

Ce projet de recherche s'effectuera dans le cadre d'une convention de collaboration réalisée entre la société Audensiel Healthcare et l'Université dans laquelle seront spécifiées les conditions de valorisation et de propriété intellectuelle.
Les analyses de données issues des simulations pourront être valorisées sous forme de publications.

Début de la thèse le 1 octobre 2023