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 Diderot : Matthieu Resche-Rigon
Responsable pour l'Université Paris Descartes : Isabelle Boutron

PROPOSITION DE SUJET DE THESE

SIGLE ET NOM DU LABORATOIRE : INSTITUT PIERRE LOUIS D'EPIDEMIOLOGIE ET DE SANTE PUBLIQUE UMR S 1136

NOM DE L'EQUIPE : MALADIES TRANSMISSIBLES : SURVEILLANCE ET MODELISATION

DIRECTEUR DE THESE : RENAUD PIARROUX

ADRESSE : LABORATOIRE DE PARASITOLOGIE-MYCOLOGIE HOPITAL DE LA PITIE SALPETRIERE 75013 PARIS

TITRE DE LA THESE : APPORT DES APPROCHES D'APPRENTISSAGE PROFOND A L'IDENTIFICATION DES AGENTS FONGIQUES IMPLIQUES EN PATHOLOGIE HUMAINE, ANIMALE ET VEGETALE PAR SPECTROMETRIE DE MASSE.

CO-ENCADRANT EVENTUEL : XAVIER TANNIER

EQUIPE ET LABORATOIRE DU CO-ENCADRANT : LABORATOIRE D'INFORMATIQUE MEDICALE ET D'INGENIERIE DES CONNAISSANCES EN E-SANTE UMRS_1142

PRESENTATION DU SUJET

1. *le contexte scientifique du projet ;*

Les infections fongiques, notamment, sont un problème de santé majeur en pathologie humaine et animale. En agriculture aussi, l'impact des espèces fongiques est très important puisque l'on estime que les agents fongiques et les oomycètes sont à l'origine de 70% des maladies des cultures. Qu'elles surviennent, chez l'homme, chez l'animal ou chez les végétaux, les maladies dues aux espèces fongiques sont particulièrement difficiles à diagnostiquer. En particulier l'identification des agents fongiques est rendue particulièrement ardue par la diversité des espèces en cause. La spectrométrie de masse par MALDI-TOF (matrix-assisted laser desorption/ionization – time of flight), une nouvelle technique diagnostique vient d'émerger et connait un essor spectaculaire en microbiologie. Cette technique, basée sur l'analyse des spectres de masse après migration dans un tube à vide, permet d'obtenir une sorte de code-barres utilisé ensuite pour identifier l'échantillon testé.

Pour faciliter l'identification des agents fongiques par spectrométrie de masse MALDI-TOF, nous avons récemment développé une application en ligne permettant de comparer les spectres obtenus à une base de données représentant plus de 1000 espèces fongiques impliquées en pathologie humaine, animale et végétale (Normand AC et al, 2017, Dupond D et al, 2018 et Imbert et al, 2019). La méthode mise en œuvre comporte une série d'algorithmes permettant de caractériser le niveau de ressemblance entre un spectre à identifier et les spectres de références connues. Cet outil a déjà été adopté par plus d'une centaine de laboratoires en France et à l'étranger. Plus de 200.000 spectres de masses ont été passés sur l'application durant l'année 2019. Les utilisateurs sont pour l'essentiel des laboratoires médicaux de grands hôpitaux, des laboratoires de recherche, des centres de référence en mycologie (humaine, animale ou végétale) ou des laboratoires médicaux ou vétérinaires privés ou public. Cependant, malgré ce succès, la méthode utilisée ne tire pas complètement parti de la signature spectrale de chaque isolat. En effet, pour des raisons de simplicité de codage, l'essentiel de

l'information tirée des spectres est résumé à une liste de quelques dizaines de pics simplement caractérisés par leur position sur le spectre. De ce fait, la méthode rencontre encore des difficultés pour l'identification d'espèces appartenant à un même complexe d'espèces et n'est pas assez précise pour détecter la présence de clones épidémiques ou pour distinguer des phylums associés à une résistance aux antifongiques. Très récemment, une première tentative d'apprentissage automatique avec des réseaux de neurones profonds, nous a permis d'obtenir des résultats prometteurs dans la détection d'isolats appartenant à un même clone épidémique au sein d'une soixantaine de souches d'*Aspergillus flavus*. Ces résultats, non encore publiés, méritent cependant d'être reproduits à plus grande échelle.

2. *les questions posées ;*

Quelle est la meilleure approche IA pour améliorer la capacité de la spectrométrie de masse à individualiser des clones ou des phylums au sein d'une espèce donnée ?

Dans quelle mesure est-il possible, en se reposant sur des méthodes d'IA de distinguer, au sein d'un ensemble de spectres de masse, ceux correspondant à un clone en cours de diffusion épidémique ou à un phylum spécifique associé à un haut niveau de résistance ?

3. *les sources de données qui seront utilisées ;*

Dans un premier temps, nous nous intéresserons à la recherche de clones particuliers au sein des espèces les plus fréquemment retrouvées (les levures du genre *Candida* : *C. albicans*, *C. glabrata*, *C. parapsilosis* et les moisissures du genre *Aspergillus*) pour lesquelles nous disposons déjà de milliers de spectres par espèce. De plus, grâce à notre réseau de laboratoires partenaires nous aurons la possibilité de générer, à partir de souches dont nous disposons déjà, autant de spectres que nécessaire pour simuler des clones épidémiques.

4. *les méthodes ;*

La tâche que nous nous assignons correspond, en première intention, au domaine de la classification de séquences en apprentissage statistique (ici, des séquences de valeurs issues des spectres de masse), sur laquelle on peut appliquer différentes architectures de réseaux de neurones à convolution ou récurrents (*Deep Learning*) ayant démontré leur efficacité pour différentes tâches (Sutskever et al, 2014). Elle peut s'apparenter à une classification multi-classe (autant de classes à prédire que d'espèces étudiées), mais peut également être assimilée à une identification de paires du même type (appartenant au même clone, au même phylum ou à la même espèce), par l'apprentissage d'une mesure de similarité entre deux isolats (pour déterminer, selon la granularité visée, s'ils appartiennent au même groupe). Il s'agit alors d'une classification binaire.

Dans le cas de la classification multi-classe, on peut s'attendre à de bons résultats à une granularité large (classification au niveau du genre) ainsi que pour les espèces très représentées dans le jeu de données, mais la fiabilité pourra diminuer pour des granularités fines (au niveau de l'espèce, puis de phylums spécifiques) ou des classes rares. L'approche de classification binaire permet de mêler les classes fréquentes et les classes peu fréquentes, avec l'hypothèse que les variations intra-espèce ont des caractéristiques communes que la machine pourra apprendre de façon globale. Cet apprentissage à base de similarité semble a priori plus pertinent pour certains cas d'étude, comme les clones ou la détection d'isolats résistants.

Dans les deux cas, la question de la représentation des spectres et du problème sera primordiale. En effet, les spectres bruts comportent des variations importantes (bruit, intensité, translation et étirement) et la question

de l'alignement et de la détermination des caractéristiques communes reste ouverte. Dès lors, les prétraitements et les choix de représentation (séquence de valeurs, liste de pics, écartements significatifs entre les pics) pourront remettre en cause l'évidence d'un modèle de classification de séquences pour orienter les recherches vers des modèles plus originaux.

Par ailleurs, du point de vue de l'apprentissage automatique, un intérêt méthodologique fort du projet est la possibilité de générer de nouveaux spectres synthétiques à partir d'espèces particulières grâce à des approches génératives de séquences (GAN), avec des variations de contraintes spécifiques, pour se focaliser sur certains aspects difficiles (par exemple, l'identification de la diversité potentielle des spectres issus d'un même clone). Enfin, d'autres approches seront à tester telles que les arbres de décision, celles-ci ont également prouvé leur efficacité en termes de classification de spectres de masses (Hummel J. et al, 2010 ; Datta S. et al, 2010).

Références:

Datta S, Pihur V. Feature selection and machine learning with mass spectrometry data. Methods Mol Biol. 2010
Hummel J. et al. Decision Tree Supported Substructure Prediction of Metabolites from GC-MS Profiles, Metabolomics, 2010.
Lantao Yu, Weinan Zhang, Jun Wang, Yong Yu, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient 2016.
Normand AC et al. Validation of a New Web Application for Identification of Fungi by Use of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol. 2017
Sutskeve et al. Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Syst. 4.

5. *puissance de l'étude/nombre de sujets ;*

Nous disposons déjà de plus de 100.000 spectres parfaitement labélisés. De plus, il existe autour de notre application un réseau de plus de 100 laboratoires qui nous permettra : 1) d'obtenir facilement d'autres spectres et de les labéliser ; 2) d'avoir une bonne représentation de la diversité des méthodes de préparation et d'obtention des spectres (réglage et calibration des spectromètres…) ; 3) de tester la capacité à identifier un clone (en faisant appel à des approches GAN, mais aussi en réalisant des sous-cultures d'une souche donnée dont les spectres seront générés par nos laboratoires partenaires) ; 4) de valider les méthodes d'IA mises au point dans des conditions de routine (puisque l'application est déjà utilisée en routine dans ces laboratoires).

6. *le calendrier prévisionnel;*

Première année : revue de la littérature sur les systèmes d'apprentissage profond, leur architecture et évaluation de leurs performances ; prise en main par le doctorant de la technologie de spectrométrie de masse MALDI-TOF et des algorithmes utilisés par l'application mise en ligne ; construction d'une première base d'apprentissage ; constitution d'un jeu de données spectrales simulant la circulation des clones épidémiques dans plusieurs centres ; collecte de clones résistants à partir de l'activité de routine des laboratoires (collecte déjà initiée) ; développement des premiers systèmes d'identification/classification des spectres basés sur des réseaux à convolution 1D et autres approches d'apprentissage supervisé.

Deuxième année : poursuite du développement des systèmes d'identification/classification pour en améliorer les performances et évaluation des résultats ; rédaction d'un premier article scientifique portant sur l'apprentissage profond appliqué à la classification de spectres de masse et son utilisation à des fins d'identification en mycologie médicale ; poursuite des expérimentations appliquées à la détection de clones au sein des espèces (détection de clones épidémiques, détection de souches résistantes) ; rédaction d'un deuxième

article portant sur les méthodes de repérage de clones au sein des espèces fongiques à partir des profils de spectrométrie de masse et les résultats obtenus.

Troisième année : mise en place des systèmes d'identification sur l'application en ligne (déjà disponible mais fonctionnant actuellement avec des algorithmes dirigés) et évaluation des résultats en situation de routine grâce à notre réseau de partenaires ; rédaction du mémoire de thèse.

7. *le thème de chacun des articles prévus. Une proposition de sujet de thèse doit comporter au moins deux articles originaux.*

Article 1 : Apprentissage profond appliqué à la classification de spectres de masse à des fins d'identification en mycologie médicale.

Article 2 : Repérage de clones/phylums particuliers à partir des profils de spectrométrie de masse : méthodes et résultats préliminaires.

PREREQUIS, FORMATION : CONNAISSANCES (NIVEAU MASTER 2) EN APPRENTISSAGE AUTOMATIQUE, RECONNAISSANCE DE FORMES, ANALYSE ET FOUILLE DE DONNEES. EXPERIENCE DANS LA CONCEPTION ET LE DEVELOPPEMENT DE SYSTEMES INTELLIGENTS. MAITRISE DES TECHNOLOGIES WEB (HTML, CSS, SQL, FRAMEWORK DJANGO).

CONTACT : RENAUD PIARROUX **EMAIL** : RENAUD.PIARROUX@APHP.FR **TELEPHONE** : 01 42 16 01 00 / 06 75 26 59 17

SPECIALITE DE LA THESE

Santé publique - Bioinformatique	X

VISA DU DIRECTEUR DU LABORATOIRE

(DEROGATION DE SIGNATURE NON ACCEPTEE)

AVIS FAVORABLE ☒

SIGNATURE