Institut Cochin

PhD opportunity on Biomedical Data Intelligence & Data mining

Titre : Cartographie et analyse phénotypique des connaissances sur les maladies auto-immunes
Title: Knowledge graphs approaches for the phenotypic analysis of autoimmune diseases

Contexte et problématique :
Les maladies auto-immunes (ou MAI) résultent d'un dysfonctionnement du système immunitaire (ou SI) qui se caractérise par une destruction des composants normaux de l'organisme par le SI Les MAI sont caractérisées par la présence d'auto-anticorps (AAC) qui sont des anticorps produits par le système immunitaire et dirigés contre une ou plusieurs protéines de l'individu lui-même.

Aujourd'hui, les connaissances sur de nombreuses maladies, même rares, sont structurées et répertoriées dans des bases de données experte bien établies, et la collecte de nouvelles données les concernant s'appuie sur l'existence de nomenclatures (ou ontologies) précises et contrôlées (comme ICD ou HPO). Cependant, peu de ces ressources permettent la récolte ou l'analyse spécifique des MAI, à contrario des maladies génétiques par exemple qui sont mieux décrites et cartographiées. Par exemple il n'existe pas de thesaurus officiel cataloguant l'ensemble des AAC associés aux MAI.

Actuellement, les connaissances de l'état de l'art sur les MAI sont disponibles dans la littérature scientifique, et de nouvelles pourraient être découvertes à partir de l'analyse de données cliniques. Cependant, la compartimentation par spécialité médicale des articles, des nomenclatures et des thésaurus actuels empêche d'avoir une vision globale des MAI. Il résulte de ce contexte qu'il n'existe aucune ressource qui liste de façon transversale les associations <MAI, signe clinique> ou <MAI, présence d'auto-anticorps>.

Projet de thèse :
La considération combinée des connaissances de l'état de l'art (la littérature scientifique, les bases de données biologiques expertes, etc.) et les entrepôts de données cliniques doivent permettre de structurer, puis d'enrichir les connaissances actuelles. En particulier nous proposons :

- de caractériser les MAI par des ensembles de signes cliniques spécifiques à une ou un groupe de pathologies
- d'associer la présence d'AAC avec une ou un groupe de pathologies, et
- d'enrichir les ontologies existantes comme HPO, pour autoriser le référencement des signes des MAI encore manquants.

Ces propriétés des pathologies devraient nous permettre d'établir des graphes sémantiques, source particulièrement intéressante pour la découverte de nouvelles connaissances dans le domaine (la proximité entre pathologies par exemple). Nous attacherons une attention particulière à tracer la provenance des connaissances extraites ou établies, à produire des sources de connaissances à la fois ouverte et FAIR (https://www.go-fair.org/fair-principles/). La comparaison de connaissances entre l'état de l'art et ce qui pourra être appris à partir des données observés cliniques constituera un des axes informatiques de la thèse.

Objectifs concrets de la thèse :
- extraction et analyse des connaissances de l'état de l'art à partir de la littérature et des bases de données expertes (notamment PubMed, OrphaNet, HPO, IGMT),
- extraction et analyse des données des entrepôts de données de santé pour valider les connaissances extraites de la littérature et les compléter par de nouvelles (EDS, BNDMR),
- construire et publier deux ressources de référence de façon ouverte et FAIR : un thésaurus des auto-anticorps, et une cartographie sous forme de graphes de connaissances qui relient pathologies auto-immunes, signes cliniques observés et présence d'auto-anticorps,
- comparaison entre les connaissances d'origines diverses.

Compétences attendues : autonomie, esprit d'équipe, compétences techniques (Python, SQL, R), expérience préalable en sciences de données ou en traitement automatique du langage appréciée, intérêt pour le domaine de la santé, excellentes compétences relationnelles et sociales.

Mots clés : maladies rares, maladie auto-immunes, auto-anticorps, extraction de connaissances, graphe de connaissances, traitement automatique des langues, fouille de données, bases de données de santé.

Où ? A Paris Santé Campus (75015 Paris) et l'Institut Cochin (75014 Paris). L'inscription sera faite dans l'Ecole Doctorale ED386 de l'Université Paris Cité.

Quand ? Date de début prévue : automne 2022

Financement ? Plusieurs demandes de bourses (deadlines : 13 mai, 18 mai, 14 juin 2022)

Comment postuler ? Envoyer un CV, une lettre de motivation d'une page (expliquant les motivations scientifiques ainsi que les projets futurs) et 1 à 2 contacts pour recommandation.

Contact :
Maud De Dieuleveult, CRCN Inserm (maud.de-dieuleveult@inserm.fr, page perso),
Adrien Coulet, CRCN Inria, HDR (adrien.coulet@inria.fr, page perso).
La thèse se fera également en collaboration avec le Dr. Anne-Sophie Jannot, MCUPH, AP-HP – UPC (annesophie.jannot@aphp.fr)