Génération de Modèles QSAR à travers KNIME®

Introduction

Cette formation de trois jours offre une étude pratique et approfondie sur la génération de modèles chémoinformatiques de classification QSAR sur l'environnement KNIME®. La formation est destinée aux personnes ayant la pratique de l'environnement KNIME® et une première expérience préalable dans l'analyse de données et la chémoinformatique. Les thèmes abordés au travers du cours sont la collecte de données chémoinformatiques depuis une base de données publique, leur nettoyage, transformation en descripteurs, la génération de modèles QSAR à partir de ces données et l'analyse des résultats statistiques. Le fils conducteur de la formation est la programmation de A à Z sur KNIME® d'un modèle de classification QSAR. Les exercices de la formation sont basés sur des données HTS pharmaceutiques issues de la base de données publique PubChem.

A qui s'adresse cette formation ?

Cette formation s'adresse aux scientifiques qui souhaitent acquérir une connaissance solide de la mise en place de modèles QSAR à partir de données HTS sur l'environnement KNIME®.

Vous allez apprendre à

Collecter des données chémoinformatiques, les nettoyer, transformer en fingerprints et descripteurs, identifier et filtrer des descripteurs, générer des modèles QSAR performants à partir de ces données, et analyser leurs résultats statistiques.

Quels sont les prérequis

•Avoir suivi les formations PIKAÏROS « Initiation à l'Exploration de Données à travers KNIME » et « Initiation à l'Analyse de Données à travers KNIME » ou avoir une expérience prouvé équivalente en analyse de données avec KNIME®

•Connaitre les bases de la chémoinformatique

•Comprendre l'anglais informatique élémentaire (outil en langue Anglaise.)

Modalités, délais d'accès et adaptation aux personnes en situation d'handicap

Cette formation est entièrement basée sur le logiciel pédagogique KNIME avec la mise en place de workflows comme méthode d'apprentissage. Une évaluation continue par QCM est effectuée sur les exercices proposés lors de la formation. La formation est proposée uniquement en présentiel en intra-entreprise. Possibilité d'adapter le temps et mode de formation aux personnes en situation d'handicap. Délais d'accès maximum d'un mois entre la commande et le début de la formation.

Programme complet

Premier jour

Collecte et préparation de données Chémoinformatiques

- Collecte et traitement des données HTS : Quelles sont les sources publiques de données disponibles pour des modèles QSAR. Quelles sont les difficultés expérimentales lors de l'utilisation de données QSAR. Comment évaluer la qualité de vos données et les préparer pour la mise en place d'un modèle QSAR.
- Génération, transformation et filtrage de données : génération et choix de variables structurales et physico-chimiques. Etude de la corrélation entre variables et sélection de variables les plus pertinentes par des méthodes rapides et force-brute.
- Pourquoi et comment diviser un ensemble de données d'apprentissage pour des modèles QSAR : Échantillonnage des données, application des techniques de validation croisée. Techniques de re-échantillonnage.

Exercices : Collecte, traitement et filtrage de données Chémoinformatiques dans l'environnement KNIME

- Lecture de données HTS de Leishmania depuis la base de données publique PUBCHEM.
- Nettoyage et normalisation des molécules.
- Génération des variables structurales et physicochimiques. Évaluation de la corrélation entre les variables générées, filtrage et sélection de variables rapide par arbre de décision et exhaustif par force-brute.
- Génération des ensembles d'Apprentissage, Validation Croisée et Test.

Deuxième jour

Modèles de classification QSAR

Etude et comparaison de différents modèles performants de classification sur des données QSAR. Comment et pourquoi choisir un modèle QSAR

- Arbres Décisionnels, Forêts d'Arbres, Xgboost
- Réseaux de Neurones Multicouches
- Ensembles de Classification.

Méthodes de rééquilibrage des données déséquilibrées de classification QSAR

Etude de l'importance de l'équilibre des classes dans la classification de données. Etude de différentes méthodes de rééquilibrage, leurs avantages et désavantages

- Sous-échantillonnage
- Sur-échantillonnage
- Adaptation des fonctions de coût lors de l'apprentissage
- Adaptation du seuil de décision des ensembles après apprentissage
- Equilibrage des données dans les ensembles de classification lors de l'apprentissage.

Exercices : Mise en place des modèles de classification QSAR sur l'environnement KNIME

- Mise en place pratique d'un modèle de référence par arbres de décision sur les données récoltées et nettoyées de Leishmania avec sous et sur-échantillonnage
- Mise en place pratique et comparaison des modèles par forêts d'arbres décisionnelles, xgboost, réseaux de neurones multicouches avec modification de fonction de coût ou de seuil de décision
- Construction d'un ensemble de classification avec rééquilibrage de données
- Test et évaluation des résultats fournis par les estimateurs statistiques.

Troisième jour

Méthodes de sélection de variable de données pour la classification QSAR

Etude de différentes méthodes de sélection de variable, avantages et désavantages

- Méthodes approximatives Backward and Forward
- Méthode exhaustive Force-Brute
- Sélection supervisée de variables par arbres de décision avec brassage des données
- Optimisation par mélange des méthodes de sélection de variables
- Interprétabilité des modèles et des variables sélectionnées.

Amélioration et méthodes d'évaluation de la qualité d'un modèle de classification QSAR

- Les estimateurs les plus performants de la qualité d'un modèle et leurs biais statistique
- Comment déterminer la qualité des modèles QSAR et équivalence statistique des modèles
- Définition et importance des domaines d'applicabilité fixe et stratifié
- Comment mettre en valeur les résultats de classification de vos modèles.

Exercices : La Modélisation de Données QSAR en pratique

- Mise en place d'un domaine de validité fixe et stratifié avec représentation cartographique
- Evaluation et comparaison pratique des différents modèles de classification mis en place
- Mise en place d'un reporting avec estimateurs statistiques et domaine de validité stratifié
- Mise en production d'un modèle QSAR validé et opérationnel.