QSAR Model Generation through KNIME®

Introduction

This three-day training offers a practical and in-depth study on the generation of QSAR classification models in the KNIME® environment. The course is intended for people with experience in the KNIME® environment and some prior experience in data analysis and cheminformatics. The topics covered throughout the course include the collection and cleaning of cheminformatics data from a public database, the generation of descriptors, the implementation of QSAR models from these data and the analysis of the statistical results. The overall objective of the training is to program a QSAR classification model from A to Z. The training exercises are based on pharmaceutical HTS data from the public database PubChem.

Who is this course for?

This training is intended for scientists who want to acquire a solid knowledge of the implementation of real and operational QSAR classification models based on HTS data in the KNIME® environment.

You will learn how to . . .

Collect cheminformatics data, verify its quality and clean it, transform it into fingerprints and descriptors, identify and filter the most informative descriptors, generate efficient QSAR classification models from these data, compare the models and analyze their statistical results.

To follow this course, you should . . .

- Have followed the PIKAÏROS training courses "Initiation to Data Exploration through KNIME" and "Initiation to Data Analysis through KNIME" or have a proven equivalent experience in data exploration and analysis with KNIME®
- Know the basics of cheminformatics
- Understand basic computer English (the KNIME® interface is in English.)

Terms, access times and adaptation to people with disabilities

This course is entirely based on the KNIME educational software with the implementation of workflows as a learning method. A continuous evaluation by MCQ is carried out on the exercises proposed during the training. This course is only offered face-to-face at the client's premises. Possibility of adapting the time and mode of training to people with disabilities.

This course should be booked at least 1 month in advance of the desired training date.

Full programme

First day

Cheminformatics data collection and preparation

- HTS data collection and processing: Which public data sources are available for QSAR models? What are the experimental difficulties when using QSAR data? How to assess the quality of your data and prepare it for the implementation of a QSAR model.
- Data generation, transformation and filtering: generation and choice of structural and physchem variables. Study of the correlation between variables and selection of the most relevant variables by rapid and brute-force methods.
- Why and how to split a training dataset for QSAR modeling: Sampling the data, applying cross-validation techniques. Resampling techniques.

Exercises: Collecting, processing and filtering of Cheminformatics data in the KNIME environment

- Reading Leishmania HTS data from the public database PubChem.
- Cleaning and standardizing molecules.
- Generating structural and physchem variables. Evaluating the correlation between the generated variables, fast filtering and selection of variables by decision trees and exhaustive filtering by brute force.
- Generating Training, Cross Validation and Test sets.

Second day

QSAR classification models

Study and comparison of different efficient classification models on QSAR data. How and why to choose a QSAR model

- Decision Trees, Tree Forests, Xgboost
- Multilayer Neural Networks
- Classification Ensembles

Methods for rebalancing unbalanced QSAR classification data

Study of the importance of class balancing in data classification.

Study of different rebalancing methods, their advantages and disadvantages:

- Down-sampling
- Over-sampling
- Adaptation of cost functions during learning
- Adaptation of the decision threshold of Ensemble Models after learning
- Balancing data in Ensemble Models during training

Exercises: Implementation of QSAR classification models in the KNIME environment

- Practical implementation of a reference model by decision trees on the collected and cleaned Leishmania data with under-sampling and over-sampling
- Practical implementation and comparison of models for instance random forests, xgboost and multilayer neural networks with modified cost functions or decision thresholds
- Construction of a classification set with data rebalancing
- Testing and evaluation of results provided by statistical estimators

Third day

Data variable selection methods for QSAR classification

Study of different variable selection methods, advantages and disadvantages:

- Heuristic Backward and Forward methods
- Exhaustive Brute-Force methods
- Supervised selection of variables by decision trees with data shuffling
- Optimization of variable selection by mixing different methods
- Interpretability of models and selected variables

Improvement and quality assessment of QSAR classification models

- The most efficient estimators of QSAR model quality and their statistical biases
- How to determine the quality of QSAR models and their statistical equivalence
- Definition and importance of fixed and stratified applicability domains
- How to highlight the classification results of your models

Exercises: QSAR Data Modeling in Practice

- Implementation of a fixed and stratified applicability domain with cartographic representation
- Evaluation and practical comparison of the different QSAR classification models implemented during the course
- Implementation of reporting components with statistical estimators and stratified applicability domain
- Production of a validated and operational QSAR model