Introduction

This three-day course gives an introduction to the fundamentals of Cheminformatics within the KNIME® environment. The training is intended for people already familiar with KNIME® and who have previous experience in data mining. The main theme addressed through the course is the implementation of cheminformatics workflows on the KNIME® environment for the use of chemical databases, curation of molecules, generation of descriptors and fingerprints, selection of diverse and focused libraries and chemical data analysis and visualization. This training is based on the existing RDKit nodes on KNIME.

Who is this course for?

The training is intended for scientists who want to learn the basis of cheminformatics for the first time on the KNIME® environment with generic practical examples to illustrate the cheminformatics principles.

You will learn how to . . .

- Access and transform public cheminformatics libraries under different molecular formats: Smiles, SMARTS, Mol, SDF, InChi, IUPAC
- Process, clean, normalize and draw chemical structures
- Generate different molecular fingerprints and PhysChem descriptors
- Generate molecular Murckos, Frameworks and Fragments
- Cluster libraries using different metrics and methods and select chemical series
- Search molecules by similarity, substructure or R-Groups
- Design diverse and focused molecular libraries.

To follow this course, you should . . .

- Have followed the PIKAÏROS training course “Introduction to Data Mining through KNIME®” (IDMK) or have equivalent in-depth experience in Data Mining in the use of the KNIME® environment.

- Understand basic computer English (the KNIME® interface and documents are 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 offered only face-to-face in intra-company at the client's premises.

Possibility of adapting the time and mode of training to people with disabilities.

Consider 1 month waiting time to access the training from the order.

Full programme

First day

Introduction to the theory and practice of Cheminformatics

The basis of Cheminformatics

- From real compounds to graphical molecular representations and file formats:
  - Smiles, SMARTS, Mol, SDF, InChi, IUPAC

Curation and standardization of Chemical Structures

- Differentiate non-organic from organic structures
- Identify and separate salts from molecules
- Check structure connectivity
- Generate Canonical Smiles and Tautomers
- Detect duplicated molecules using InChi and complementary techniques

Access and Manipulation of a Cheminformatics Database

- Main Cheminformatics public databases such as ChEMBL and PubChem
- What is a cheminformatics database, a scheme and its possible formats
- How to store and retrieve important information from a cheminformatics database
- How to relate compounds and targets in a cheminformatics database

Exercises: First retrieval and manipulation of cheminformatics datasets

- Access, analysis of a public database (PubChem or ChEMBL)
- Retrieval of specific cheminformatics data given an HTS target from a public database
- Curation of previous set of molecular structures
- Storage under different cheminformatics formats

Second day

Generation of new information from chemical structures

Fingerprints

- Definition and usage of fingerprints
- Principal families of fingerprints: Linear, radial, structural, others
- Fingerprints available through KNIME nodes: MACCS, Morgan, FeatMorgan, AtomPair, etc.
- How to generate, manipulate and transform fingerprints with KNIME

PhysChem Descriptors

- Definition and usage of physchem properties
- Principal families of physchem properties
- How to generate, manipulate and transform physchem properties with KNIME

Fingerprint and PhysChem Descriptor selection based on target data

- Correlation / Decorrelation of fingerprints and descriptors (variables)
- Variable selection by variance and correlation
- Interpretation and visualization of selected variables

Exercises: Generation, conversion and manipulation of fingerprints

- Generation and comparison of most common fingerprints from previously curated molecular datasets
- Generation of simple Decision Tree classification models with and without variable selection

Third day

Generation and clustering of diverse, dense and focused molecular libraries

Molecular Clustering

- Definition of molecular clustering and use-cases
- Different metrics, strategies and nodes to cluster molecular datasets

Diversity, Similarity & Density

- Definition and use-cases of molecular Diversity, Similarity and Density
- Different methodologies and nodes to implement Diversity, Similarity and Density

Structural & PhysChem Chemical Series

- Definition and use-cases of Chemical Series in Cheminformatics
- Different methodologies and nodes to implement Chemical Series based on structure, Murcko and fragments
- How to build and organize a chemical collection from scratch

Exercises: Practical molecular clustering based on different metrics, strategies and objectives

- Clustering and comparison of previously curated collections based on different fingerprints, descriptors and substructure representations
- Projection of small curated datasets on big clustered chemical libraries
- Selection of sub-ensembles of libraries based on similarity and diversity
- Search of molecules in big collections by similarity and substructure