Predictive global models of cruzain inhibitors with large chemical coverage

Jose Guadalupe Rosas-Jimenez,[†,¶] Marco A. Garcia-Revilla,[†] Abraham Madariaga-Mazon,[‡] and Karina Martinez-Mayorga[*,‡]

[†]*Division de Ciencias Naturales y Exactas, Universidad de Guanajuato, Guanajuato, Mexico*

[‡]*Instituto de Quimica, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico*

[¶]*Current address: Instituto de Quimica, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico*

E-mail: kmtzm@unam.mx

Python script for the calculation of pIC50 values of cruzain inhibition, using K Nearest Neighbors and Random Forest regression algorithms. Details on modeling and validation can be found in the cited paper.

The script needs the following packages:

rdkit https://github.com/rdkit/rdkit

mordred https://mordred-descriptor.github.io/documentation/master/introduction.html

sklearn https://scikit-learn.org/stable/install.html

numpy https://numpy.org/install/

pandas https://pandas.pydata.org/getting_started.html

If you do not have the above packages, please follow the installation instructions provided. It is recommended to use the Anaconda python distribution since most of the required modules are included by default: https://www.anaconda.com/products/individual

The usage of this script is:

python qsar_cruzain_predictor.py [-h] [-i input.sdf] [-o output.csv]

-h, Print help message

-i, SDF file with 2D structures of molecules to be predicted

-o, name of the csv file where results will be saved

The default value for the csv with the results is output.csv. If no sdf is provided as input, the program will print a help message. The sdf file just needs the 2D coordinates of the molecules. The script takes the molecule names stored in the sdf files and uses it as IDs in the results (see below).

The output of the script is a csv file with the following headers:

First 18 columns: Values of the mordred molecular descriptors for each molecule (non scaled).

pIC50_KNN: pIC50 values predicted by the KNN algorithm.

pIC50_RF: pIC50 values predicted by the Random Forest algorithm.

Max Similarity: Maximum MACCSKeys based Tanimoto similarity of the molecule against molecules in the training set.

Leverage KNN: Leverage values in the KNN feature space.

Max Leverage KNN: Maximum value of leverage to be considered as inside applicability domain.

Leverage RF: Leverage values in the Random Forest feature space.

Max Leverage RF: Maximum value if leverage to be considered as inside applicability domain.

ID: Molecule identifier (molecule name in the SDF file)

This folder contains the following (required) files:

qsar_cruzain_predictor.py: Main python script to make the predictions using the trained algorithms.

utilities.py: File with some classes to use as tools for applicability domain definition.

utils/AD_KNN.pickle: Pickle file with information about the training set to calculate leverage of KNN feature space.

utils/AD_RF.pickle: Pickle file with information about the training set to calculate leverage of RF feature space.

utils/ADsimilarity.pickle: Pickle file with information about the the training set to calculate similarity.

utils/KNN_CRUZ.pickle: Pickle file with the scaling and KNN trained model.

utils/RF_CRUZ.pickle: Pickle file with the scaling and RF trained model.