Applicability Domain using Standardization approach

13 April 2016

DTC Laboratory
Dept. of Pharm. Tech.,
Jadavpur University,
Kolkata, West Bengal,
INDIA

AD using standardization approach

Applicability Domain (AD) is simply defined as "the response and chemical structure space in which the QSAR model makes predictions with a given reliability".

Recently our group has proposed a new and simple approach [1] to find the applicability domain. The "AD using Standardization approach" is a tool to find out compounds (test set/query compounds) that are outside the applicability domain and also detect outliers from training set compounds.

Basic principle behind this approach:

The basic principle applied in this approach is as follows:

A QSAR model learns from the features of the training set compounds. The developed model is then applied for prediction of test set compounds which should be structurally similar to the training set compounds so that the model can perform well based on the similarity principles. If a small fraction of the training set contains features very dissimilar to the rest and most of the compounds, then obviously those features are not properly included in the training process. These compounds are *X*-outliers. If test set compounds are similar to these small fraction of training set compounds, then their predictions are expected not to be good, as the model has not captured the features of those training set compounds which have a small representation and are different from majority of the compounds. So those test set compounds are expected to be outside the AD of the model. Again, the test set compounds which are not similar to any of the training set compounds are also outside the AD. Ideally, all the

descriptors of the training set compounds should follow a normal distribution pattern. According to this distribution, 99.7% of the population will remain within the range *mean ±3 standard deviation (SD).* Thus, *mean ±3SD* represents the zone where most of the training set compounds belong to. Any compound outside this zone is dissimilar to the rest and majority of the compounds. Thus, after a descriptor column is standardized based on the corresponding mean and standard deviation for the training set compounds only, if the corresponding standardized value for descriptor *i* of compound *k* (S_{ki}) is more than 3, then the compound should be an *X*-outlier (if in the training set) or outside AD (if in the test set) based on descriptor *i*. This test should run for all descriptors present in the model (however, we have not considered, for the sake of simplicity, inter-correlation among descriptors and relative importance of the descriptors to the model, which are the drawbacks of this method). If the maximum S_i value of a compound *k* is lower than 3, then the compound is quite similar to a good number of compounds in the training set with respect to all descriptors (not an *X*-outlier if in the training set and is within AD if in the test set). If the minimum S_i value of a compound k is higher than 3, then the compound is quite dissimilar to most of the compounds in the training set with respect to all descriptors (an *X*-outlier if in the training set and not within AD if in the test set). If the compound has a maximum S_i value above 3 but the minimum S_i value is below 3, then the compound is similar to most of the training set compounds with respect to some descriptors and at the same time dissimilar to most of the training set compounds with respect to other descriptors. Thus, we need to formulate some criterion of assessment of *X*-outliers or applicability domain behavior of such compounds. Now again considering an ideal case of standardized normal distribution, the

standard score (*Z*) corresponding to 1.28 represents a relative frequency of occurrence of less than 1.28*SD being 90%. Thus, in our case, if mean of the S_i values of a compound for all descriptors in a model plus 1.28 times corresponding standard deviation (call it S_{new}) is lower than 3, then there is 90% probability that the S_i values of that compound are lower than 3. Thus, when S_{new} value of a compound is lower than 3, then the compound can be considered to be not an X-outlier (if in the training set) or within the AD (if in the test set). This assumption is statistically more valid when a higher number of descriptors are present in the model.

AD Program folder

The program folder consists of three folders "Data", "Lib" and "Output". For convenience, user may keep input file in "Data" folder and may save output files in "Output" folder, since by default, clicking on the browse button will open these folders. "Lib" folder consists of library files required for running the program. Hence try not to move or delete or rename these library files.

Input file format

Three different file types are allowed *i.e. xlsx , xls* and *csv* as input file. The input file should consist of serial number column (*first column*), and descriptor values columns (*subsequent columns*) for each object/compound. The format in which this information should be placed in the input file is as follows:

First Row: Header *i.e.* name for each column, for instances, descriptor names. *It can be numerical, alphabet or alphanumerical in nature.*

First column: Serial number/ Compound number. *It should be numerical and not alphabet or alphanumerical in nature (except first row).*

Subsequent columns: Independent variables/Descriptor values; each column will consist of descriptor values for all the compounds/objects. *These values should be numerical values and not alphabets or alphanumerical values (except first row).*

Note: No more information should be present. And, no blank cells should be present in-between. *Sample input files are provided in 'data' folder*

How to run the program

It's simple! Just click/double click on the jar file (AD.*jar*) present in the '*AD*' program folder. A window will open as shown in *Snapshot*, with few queries, which a user has to fill before clicking on '*Start*' button to run the program.

"*Select Training set File*": Click on 'browse' button to select the training set file. By default, it will open the "Data" folder present in '*AD*' program folder. So for convenience, user can keep the input files in the "Data" folder.

"*Select Test set File*": Click on 'browse' button to select the test set file. By default, it will also open the "Data" folder present in '*AD*' program folder.

"*Select Output Directory*": Click on 'browse' button to select the destination/output file directory and define output file name. By default, it will open the "*Output*" folder present in '*AD*' program folder. User can save the output files in the "Output" folder.

Output

The output consist of one file *i.e.* a excel file with two spread sheets. The content of this file is discussed below.

Note: *If the input file type is .csv, then two output files (.csv) will be generated, since multiple spreadsheets are not allowed in .csv file.*

Output _.xlsx file : Two excel sheet will be generated within one excel file (*if input file type is .xlsx or .xls*) each for training and test set. The generated excel sheet (*.xlsx/xls/csv*) will comprise of serial number column and descriptor columns from the input file. One additional column (*last column*) containing *Applicability domain (test set)/Outlier (training set) information* will be present in the respective sheets.

Reference:

1. Roy, Kunal, Supratik Kar, and Pravin Ambure. "On a simple approach for determining applicability domain of QSAR models." *Chemometr Intell Lab Syst.* 145 (2015): 22-29.

Java External Library Used

Apache POI – the Java API for Microsoft Documents

- Available at http://poi.apache.org/

XMLBeans

- Available at http://xmlbeans.apache.org/

Contact us at the following address:

Dr. Kunal Roy,

Drug Theoretics and Cheminformatics Lab.,

Dept. of Pharmaceutical Technology,

Jadavpur University,

Kolkata, West Bengal,

INDIA-700032

Email Id: kunalroy_in@yahoo.com

Software Developer details:

Pravin Ambure,

Research Scholar,

Drug Theoretics and Cheminformatics Lab.,

Dept. of Pharmaceutical Technology,

Jadavpur University,

Kolkata, West Bengal,

INDIA-700032

E-mail Id: ambure.pharmait@gmail.com