Tutorial on QSAR/QSPR modeling

Alexandre Varnek

Software: CODESSA-PRO ISIDA-QSPR, ISIDA EdChems, ChemAxon Standardizer

Datasets: ALKAN.SDF, part_coef.SDF, DataCuration.SDF

Part 1. QSAR/QSPR modeling with CODESSA-PRO (molecular descriptors).

1. Create storage.
2. Read the ***ALKAN.SDF*** file.
3. Visualize the pattern matrix.
4. Select *critical temperature* as a property.
5. Build models containing from 2 to 6 descriptors on the entire set (BMLR option).
6. Treatment of results.
   - Build plots R^2 and Q^2 *vs* descriptors number (N_d). What is an optimal N_d value?
   - Copy output of the program into a WORD file. What descriptors have been selected ?
7. Prepare a test set selecting each 5th compound starting with the 1st one. Select an ensemble of remaining compounds as a training set.
8. Repeat steps 5 and 6. Does the program select the same descriptors as in previous exercise ?
9. Apply the "best" model to the test set and calculate statistical parameters of the linear correlation *PRED vs EXP*. Compare them to those calculated for the training set.
10. Select another test set taking each 5th compound starting with the 3d one. Repeat steps 8 and 9. Do predictive performance of the model varies as a function of the test set ?

Part 2. QSAR/QSPR modeling with ISIDA (fragment descriptors).

1. Open *Training Set*
2. Read ALKAN.SDF file
3. Select *critical temperature* as a property.
4. Prepare a MASK file (no test set).
5. Select sequences of atoms and bonds as descriptors.
6. Run calculations on the entire set. Notice statistical parameters.
7. Prepare another MASK file selecting to the test set each 5th compound starting with the 1st one.
8. Run calculations and compare statistical parameters obtained for the training and test sets.
9. Select another test set taking each 5th compound starting with the 3d one. Repeat steps 8 and 9. Do predictive performance of the model varies as a function of the test set ?

10. Select augmented atoms (*atoms and bonds*) as descriptors. Repeat step 8 using the MASK file obtained at the step 7. How does predictive performance change as a function of descriptors ?
11. Perform BATCH calculations using external 5-fold cross-validation. Visualize plot R^2(test) as a function of Q^2.
12. Perform interactive modeling of new structures using *ISIDA EdChems* program.

Part 3. **Data cleaning**

Using *ChemAxon Standardizer,* clean *DataCuration.SDF* file.

Use the following options:

- Aromatize
- Clean 2D
- Clear Stereo
- Remove Fragment
- Transform

Individual exercises.

Perform QSAR modeling of partition coefficients tissue-air using CODESSA-PRO and ISIDA. read input data from the *part_coef.SDF* file.