CMF – R Package Implementing the Continuous Molecular Fields Approach

1. Inrtoduction

The package CMF contains a set of R functions that implement the Continuous Molecular Fields (CMF) approach to building 3D-QSAR models. The reference version of the R environment for statistical computing and graphics is 3.0.1.

2. Installation and Setup of the Required Software

The latest version of the R environment can be downloaded and installed from the homepage of the R Project for Statistical Computing http://www.r-project.org/. In order to perform visualization of molecules, molecular fields and fields of regression coefficients, it is necessary to install two additional packages: rgl and misc3d. They can be installed from the CRAN homepage http://cran.r-project.org/. In order to use quantum chemical molecular fields, it is necessary to install the MOPAC12 program http://openmopac.net/MOPAC2012.html. In order to run the software implementing the CMF approach to building 3D-QSAR models, it is necessary to keep all files, including all R scripts and all data files, in the same folder (directory), which should be specified as the current working directory.

4. The Basic Workflow for Building 3D-QSAR Models Using the CMF Approach

Two major modes of building and analyzing the performance of 3D-QSAR models are provided. The first one is based on a single splitting between a training set and an external (independent) test set. *The model built on the training set is further applied to predict activity values of compounds contained in the test set.* This mode requires the use of the following four input files:

- a file (in the mol2 format) containing 3D structures of the molecules belonging to the training set (its default name is ligands-train.mol2);
- a file (in the delimited txt format) containing experimental (measured) activity (property) values of the corresponding chemical compounds belonging to the training set (its default name is activity-train.txt);
- a file (in the mol2 format) containing 3D structures of the molecules belonging to the test set (its default name is ligands-pred.mol2);
- a file (in the delimited txt format) containing experimental (measured) activity (property) values of the corresponding chemical compounds belonging to the test set (its default name is activity-pred.txt).

The second mode is based on the procedure of external n-fold cross-validation. In this case the whole set of compounds is split into the training and test sets n times, so each compound appears in a test set exactly once. *The 3D-QSAR model built on a training set is applied to the corresponding test set, and all prediction results are accumulated.* This mode requires the use of the following two input files:

- a file (in the mol2 format) containing 3D structures of the molecules belonging to the whole set (its default name is ligands-all.mol2);
- a file (in the delimited txt format) containing experimental (measured) activity (property) values of the corresponding chemical compounds belonging to the whole set (its default name is activity-all.txt)

4.1. Aligning molecules

Molecules in a dataset can be aligned using two different approaches. If the set of compounds under study is congeneric, then it can be aligned by least-square fitting (algorithm arun) to a common template substructure,

which should be contained in all molecules belonging to this set. Otherwise, alignment can be performed using the seal algorithm. If the first approach is chosen, it is first necessary to obtain the template substructure to be used for performing alignment. For example, it can be extracted from some molecule by specifying a list of serial numbers of atoms. This can be accomplished using the script cmf-do-make-template.R, in which the values of the following parameters can be specified:

- mdb-fname – file name containing the structure from which the template substructure is to be extracted;
- imol – the serial number of the molecule in the file mdb-fname, from which the template substructure is to be extracted;
- atom_list – the list of serial numbers of atoms used for extracting the template substructure;
- template_fname – file name for the template substructure.

The template substructure can further be used for performing molecular alignment. This can be carried out using the following script:

```
# File name of molecular databased to be aligned
mdb_fname <- "ligands-train.mol2"

# Molecules from mdb_fname to be aligned
iimol <- c(1:72)

# File name of template
templ_fname <- "ligands-template.mol2"

# File name for aligned database
mdb_a_fname <- "ligands-aligned.mol2"

# Algorithm (arun/seal)
algorithm <- "arun"

mdb <- read_mol2(mdb_fname)
mdb_tmp <- read_mol2(templ_fname)
templ <- mdb_tmp[[1]]
if (algorithm == "arun") {
  mdb_a <- align_mdb_template(mdb, templ)
#  mdb_a <- align_mdb_template(mdb, templ, iimol)
} else if (algorithm == "seal") {
  mdb_a <- align_mdb_seal(mdb, templ)
} else {
  cat("Unknown algorithm\n")
}
write_mol2(mdb_a, mdb_a_fname)
```

in which the following parameters can be specified:

- mdb_fname – file name of the molecular database to be aligned;
- iimol – List of molecules from mdb_fname to be aligned (if this parameter is dropped, the whole molecular database is to be aligned);
- templ_fname – file name of the template substructure;
- mdb_a_fname – file name of the produced aligned database;
- algorithm – alignment algorithm (in this case, arun).

4.2. Computing kernels for the training set

The first step in building 3D-QSAR models using the CMF approach is to compute kernels for all pairs of compounds from the training set and to write them to a file. The kernels can be computed for five types of

molecular fields:

- q - electrostatic molecular field;
- vdw - steric molecular field;
- logp - hydrophobicity field;
- abra - hydrogen-bond acidity field;
- abrb - hydrogen-bond basicity field.

All kernels are computed for the following fixed set of attenuation parameter: 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1.0, 2.0, 3.0, 5.0, 10.0. To carry out this kernel computation procedure, one can run the computational R script:

```
# Training set file name
train_fname <- "ligands-train.mol2"

# The name of the file with kernels for training
kernels_train_fname <- "ligands-kernels-train.RData"

# Molecular fields
mfields <- c("q","vdw","logp","abra","abrb")

# Verbose computation of kernels (TRUE/FALSE, 1/0)
print_comp_kernels <- TRUE

comp_kernels_train(
  train_fname = train_fname,
  kernels_train_fname = kernels_train_fname,
  mfields = mfields,
  print_comp_kernels = print_comp_kernels
)
```

The values of the following parameters may be specified by editing their default values in the script:

- train_fname - the name of the mol2 file containing the structures from the training set (default: “ligands-train.mol2”);
- kernels_train_fname - the name of file with computed kernels (default: “ligands-kernels-train.RData”);
- mfields – the list of molecular fields for which kernels should be computed (default: c(“q”,“vdw”,“logp”,“abra”,“abrb”));