Command-Line Interface Example

The following section gives an example of using the binary executable version of jCompoundMapper. This version can be used in, e.g., shell scripts. Calling the command-line tool using `-h` gives an overview of possible parameters.

```
java -jar jCMapper.jar -h
usage: jCMapper
 -a     Atom Type: CDK_ATOM_TYPES: 0, ELEMENT_NEIGHBOR: 1,
        ELEMENT_NEIGHBOR_RING: 2, ELEMENT_SYMBOL: 3, CUSTOM: 4,
        DAYLIGHT_INVARIANT: 5, DAYLIGHT_INVARIANT_RING: 6
 -c     Fingerprinting algorithm: DFS: 0, ASP: 1, AP2D: 2, AT2D: 3, AP3D:
        4, AT3D: 5, CATS2D: 6, CATS3D: 7, PHAP2POINT2D: 8, PHAP3POINT2D: 9,
        PHAP2POINT3D: 10, PHAP3POINT3D: 11, ECFP: 12, LSTAR: 13, SHED: 14, RAD2D:
        15, RAD3D: 16
 -d     Distance cutoff / search depth
 -f     MDL SD file
 -ff    Output format: LIBSVM_SPARSE: 0, LIBSVM_MATRIX: 1, FULL_CSV: 2,
        STRING_PATTERNS: 3, WEKA_HASHED: 4
 -h     Print help
 -hs    Hash space size (default=1024)
 -l     Label (MDL SD Property)
 -lt    Label threshold
 -m     Distance measure (matrix format): TANIMOTO: 0, MINMAX: 1
 -o     Output file
 -s     Scaling factor (3D fingerprints)
```

Using the defaults (or via `-ff 0`), jCompoundMapper generates a hashed LIBSVM output format using the depth-first search encoding with element plus neighbor count atom types.

In the following, we process the training and the known test set from the environmental toxicity challenge (`http://www.cadaster.eu/node/65`) which were converted to MDL SD format. The label (MDL property) to be learned is `log(IGC50-1)`. Using this settings, the structures of the training set were mapped to hashed fingerprints.

```
java -jar jCMapper.jar -f challenge_train.sdf -l "log(IGC50-1)"
Processing MDL SD file: challenge_train.sdf
Selected label: log(IGC50-1)
Output format: LIBSVM_SPARSE
Fingerprinting algorithm: DFS
Search depth: 8
Labeling Algorithm: ELEMENT_NEIGHBOR
Export option: LIBSVM_SPARSE
Hash space size = 1024

Output file = challenge_train.DFS.LIBSVM_SPARSE
Time elapsed: 4196 ms
Avg. features per mol = 46.509
```

After the computation, an overall statistic is plotted showing e.g. the average number of features in the fingerprints. In the next step, we map the test file to the same representation. Bits in the test file will be set in exactly the same positions in the vector because the random numbers are generated by using the seed value defined by the features.

```
java -jar jCMapper.jar -f challenge_test_known.sdf -l "log(IGC50-1)"
Processing MDL SD file: challenge_test_known.sdf
Selected label: log(IGC50-1)
Output format: LIBSVM_SPARSE
Fingerprinting algorithm: DFS
Search depth: 8
Labeling Algorithm: ELEMENT_NEIGHBOR
Export option: LIBSVM_SPARSE
Hash space size = 1024

Output file = challenge_test_known.DFS.LIBSVM_SPARSE
Time elapsed: 3416 ms
Avg. features per mol = 54.771
```

In the next step, a cross-validation is conducted by using the precompiled binary distribution of LIBSVM [1] that can be downloaded from the LIBSVM homepage. The parameters are set as follows: `-t 0` sets the linear kernel (dot product), `-s 3` sets ϵ regression, and `-c 2` sets the error weight to 2.

```
svmtrain -t 0 -s 3 -v 10 -c 2 challenge_train.DFS.LIBSVM_SPARSE
```

LIBSVM produces no model in cross-validation mode. However, The LIBSVM cross-validations statistics shows that the model has an MSE of 0.32 and an Q^2 of 0.71, indicating a reasonable parametrization.

```
Cross Validation Mean squared error = 0.324891

Cross Validation Squared correlation coefficient = 0.712412
```

Finally, the model is trained by omitting the cross-validation flag `-v`.

```
svmtrain -t 0 -s 3 -c 2 challenge_train.DFS.LIBSVM_SPARSE
```

This step produces a separate model file, which can be used to predict the external test set. This is conducted by calling `svmpredict`.

```
svmpredict challenge_test_known.DFS.LIBSVM_SPARSE challenge_train.DFS.LIBSVM_SPARSE.model result
```

The results are printed by LIBSVM highlighting that the performance on the external test set is $MSE = 0.29$ and $R^2 = 0.74$. The result on the known test of the environmental toxicity prediction challenge would be in the top ranks of the competition.

```
Mean squared error = 0.291011 (regression)

Squared correlation coefficient = 0.742283 (regression)
```