ELUDE (1.0) Documentation

Luminita Moruz
Center for Biomembrane Research
Stockholm University, Stockholm, Sweden
e-mail: lumi@sbc.su.se

May 19, 2010

Contents

1 Introduction **2**

2 File formats **2**

3 Train a new model **2**
3.1 Running ELUDE; command line options available 2
3.2 Example 3

4 Use a pre-trained model from the library **3**
4.1 Running ELUDE; command line options available 3
4.2 Example 4

1 Introduction

ELUDE is a software package designed for predicting peptides' retention time in liquid chromatography. The core predictor in ELUDE employs ϵ-support vector regression to build an accurate model that can be then applied to estimate retention time for peptides generated under the same chromatographic conditions. In order to train an accurate model, ELUDE requires around 500 training peptides, i.e. peptides with known retention time. When only a small number of such peptides available (e.g. 50-100 peptides), as it is often the case in targeted proteomics, ELUDE implements an alternative workflow where a model is selected from a library of predictors and calibrated for the condition at hand. The remaining of this document is organized as follows. Section 2 describes the file formats used by ELUDE; Section 3 illustrates how to train a new predictor in ELUDE and Section 4 shows the use of ELUDE when little amounts of training data available.

2 File formats

A standard run of ELUDE requires two input files: one including peptides with known retention time used to train/calibrate a model and another one including the test peptides, *i.e.* the peptides for which we would like to estimate the retention times. The format of these files is straightforward. The train/calibration file should include on each line a peptide sequence and its observed retention time separated by space. The file including the test peptides should include one peptide sequence per line. If the observed retention times of the test peptides are known, then the test file should have the same format as the training file, with each line including the peptide sequence and the retention time, space-separated. In this case ELUDE will provide a summary of the quality of the predictions, including correlation between predicted and observed retention times and the size of the window around the predicted time that will contain the correct retention time in 95% of the cases. Note that the peptide can be specified either as a sequence of letters following the standard notation of the amino acids, or in the format $A.X_1...X_n.B$, where A and B represent the previous and next amino acids in the protein sequence, respectively, and $X_1...X_n$ denotes the sequence of the peptide. If the peptide is at the beginning (end) of the protein, then A (B) will be replaced with the symbol '-'. Please refer to the directory `data/` for examples of input files for ELUDE.

3 Train a new model

3.1 Running ELUDE; command line options available

The easiest way to run ELUDE is to type the following command in a terminal window (Linux) or command prompt window (Windows)[1]

```
elude -t train_file -e test_file -o out_file
```

where `train_file` should be replaced with the name of the file including the training peptides, `test_file` with the file including the peptides for which we want to predict retention time, and `out_file` with the name of the output file. The format of the training and test files is described in the previous section, while the output file will be created by ELUDE.

In addition, other options can be used to further customize the execution of ELUDE. The full list of command line options supported is displayed below.

[1]Note that in Windows this command has to be run in the directory where the file `elude.exe` is located. To navigate to this directory, go to `Start -> Run`, type `cmd` followed by enter, then use `cd folder_name` to get to the folder including `elude.exe`

`-h`	Displays a brief help including all the available options
`-v <level>`	Verbosity of output, possible values are any interger between 1 and 5. For 1 no processing information will be displayed, for 5 all information is displayed. The default value is 4
`-t <filename>`	File including the training data
`-e <filename>`	File including the test data
`-s <filename>`	File to save the retention model
`-m <filename>`	File including a previously trained model that will be loaded
`-o <filename>`	Output file. Final predictions will be dumped in this file
`-u`	All redundant peptides from the test set will be removed. Note that redundant peptides are always removed from the training peptides
`-i <filename>`	The file where in-source fragmented peptides will be saved. In-source fragmented peptides are automatically detected and removed from the train data
`-z <value>`	The enzyme used for digestion. The possible values are `NO_ENZYME, TRYPSIN, CHYMOTRYPSIN, ELASTASE`. By default `TRYPSIN`
`-x`	All non enzymatic peptides should be removed from both training and test set. For this option the peptides need to be given in the format `A.XX...XX.B` described in Section 2
`-g <filename>`	The file where the retention index will be saved. Note that ELUDE trains a new retention index for each condition at hand
`-d`	The current model will be appended to the library

3.2 Example

The folder `data/train_data/` includes examples of input files for ELUDE. If we want to train a model using the peptides from the file `data/train_data/train.txt` and then use this model to predict the retention time of the peptides in `data/train_data/test.txt` , then we could run ELUDE as follows:

```
elude -t data/train_data/train.txt -e data/train_data/test.txt -o predictions.txt
-s model.txt -i in_source_fragments.txt -g retention_index.txt -u
```

In this case the file `predictions.txt` will include the predicted retention times for the peptides in `test.txt`, the file `in_source_fragments.txt` will include the in-source fragmented peptides detected in the training set, while `retention_index.txt` will include the retention index trained by ELUDE. The option `-u` indicates that all the redundant peptides in the test set will be removed. The trained model is saved to the file `model.txt`. This model can be subsequently used to predict retention as shown below:

```
elude -m model.txt -e data/train_data/test.txt -o predictions_model.txt -u
```

Since we use a previously trained model, no training data is required in this case. The file `predictions_model.txt` will include the predicted retention time for the peptides in `test.txt`. Note however that this alternative will yield accurate predictions only when the model was trained on data from the same chromatographic conditions as the test peptides.

4 Use a pre-trained model from the library

4.1 Running ELUDE; command line options available

When only a small set of peptides with known retention times is available, it is recommended to use a pre-trained model from the internal library maintained by ELUDE. Our package implements a procedure to automatically select the library model that best fits the training data, calibrate this model and then apply it to predict retention time for the peptides of interest. This procedure can be run by issuing a command as shown below:

```
elude -t calibration_file.txt -e test_file.txt -o out_file.txt -a
```

where `calibration_file` is the file including the training peptides, `test_file` is the file including the peptides for which retention times should be predicted and `out_file` is the output file. The formats of the calibration and test files are explained in Section 2, while the output file is created automatically by ELUDE. The option `-a` indicates that the model will be selected from the library. In addition, other options can be used to further customize the execution of ELUDE. The full list of command line options supported is displayed below.

`-a`	Indicates that the model should be selected automatically from the library
`-h`	Displays a brief help including all the available options
`-v <level>`	Verbosity of output, possible values are any interger between 1 and 5. For 1 no processing information will be displayed, for 5 all information is displayed. The default value is 4
`-t <filename>`	File including the calibration peptides
`-e <filename>`	File including the test data
`-o <filename>`	Output file. Final predictions will be dumped in this file
`-u`	All redundant peptides from the test set will be removed. Note that redundant peptides are always removed from the calibration peptides
`-z <value>`	The enzyme used for digestion. The possible values are `NO_ENZYME, TRYPSIN, CHYMOTRYPSIN, ELASTASE`. By default `TRYPSIN`
`-x`	All non enzymatic peptides should be removed from both training and test set. For this option the peptides need to be given in the format `A.XX...XX.B` described in Section 2
`-b`	The fraction of the training peptides used by ELUDE to calibrate the selected model. By default 0.95 (95% of the peptides are used for calibration)

4.2 Example

The folder `data/calibrate_data/` includes examples of input files for ELUDE. If we would like to use the peptides from the file `data/calibrate_data/calibrate.txt` to select and calibrate a model from the library, and then use this model to predict retention time for the peptides in the file `data/calibrate_data/test.txt`, then we could run ELUDE with the following command:

```
elude -t data/calibrate_data/calibrate.txt -e data/calibrate_data/test.txt
-o predictions.txt -a -u
```

The predicted retention times will be saved in the file `predictions.txt`. The option `-a` indicates that the model will be selected automatically from the library and `-u` that only unique peptides should be kept in the test file.