**NVDT**

A new computational method for protein-protein interaction and non-interaction predictions from the gene sequence

**Part 1：Datasets**

**---Positive and negative samples of each species.**

Each .xlsx file has two columns and several rows. Each row represents one protein pair (positive samples are interacting protein pairs, negative samples are non-interacting protein pairs)

(1) **Real-Datasets:** Positive samples are collected from the DIP Database, and negative samples are collected from the Negatome Database. There are two species here, H. sapiens and M. musculus.

The name of the positive samples dataset file should be the name of your species followed by '\_Positive\_Real' (e.g., M. musculus\_Positive\_Real.xlsx).

The name of the negative samples dataset file should be the name of your species followed by '\_ Negative\_Real' (e.g., M. musculus\_Negative\_Real.xlsx).

(2) **Constructed-Datasets:** Positive samples are collected from the DIP Database, and negative samples are constructed by pairing the proteins located in different subcellular positions.

The name of the positive samples dataset file should be the name of your species followed by '\_Positive\_Constructed' (e.g., S.cerevisiae\_Positive\_Constructed.xlsx).

The name of the negative samples dataset file should be the name of your species followed by '\_ Negative\_Constructed' (e.g., S.cerevisiae\_Negative\_Constructed.xlsx).

**Part 2：Feature extraction and standardization**

**---** **Sequence features are extracted and further standardized from protein pairs.**

Run file 'Feature\_extraction.m' to extract the features of protein pairs in the training dataset and testing dataset.

Run file 'Feature\_standardize.ipynb' to standardize the protein pairs' feature vectors in the training dataset and the testing dataset.

**Part 3：Training datasets and testing datasets**

**---** **Files that store the features and labels of protein pairs in the training dataset and the testing dataset of each species.**

Necessary input files before running the command:

**(1) Input files of random forest classifier:** It is in .xlsx format. (This only works on Real-Dataset.)

There are four sub-tables in the Excel table, which store the training dataset feature vectors and corresponding labels, the testing dataset feature vectors and labels, respectively.

The name of this file should be the name of your species and data type followed by '\_RFInput' (e.g., H. sapien\_Real\_RFInput.xlsx and H. sapien\_Constructed\_RFInput.xlsx).

**(2) Input files of support vector machine classifier:** It is in txt format. (This works on Real-Dataset and Constructed -dataset.)

-Training dataset: The name of this file should be the name of your species followed by '\_Train' (e.g., S.cerevisiae\_Train.txt).

Suppose that the training dataset has M samples, and each sample has N-dimensional features. The form of the document is as follows:

[label 1] [1: feature 1] [2: feature 2] … [N: feature N]

[label 2] [1: feature 1] [2: feature 2] … [N: feature N]

…

[label M] [1: feature 1] [2: feature 2] … [N: feature N]

The .txt file is generated by a file named ' FormatDataLibsvm.xls. '

-Testing dataset: Same as the Training dataset. The name of this file is the same as species followed by '\_Test' (e.g., S.cerevisiae\_Test.txt).

Similar to Training data, you can make the .txt file using FormatDataLibsvm.xls

**Part 4：Training model and prediction**

**(1) random forest classifier:** Run file ‘RF\_classifier.ipynb’.

**(2) support vector machine classifier:**

Install LIBSVM. ( LIBSVM download address: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/> )

svm-train.exe: input training dataset to generate a model.

Usage: svm-train.exe [options] training\_ dataset\_ file [model\_ file]

svm-predict.exe: used to predict results of the testing dataset

Usage: svm-predict.exe [options] test\_ file model\_ file output\_ file

**Part 5：Model evaluation**

**(1) Summary of prediction results of random forest classifier**

File ‘RF\_classifier.ipynb’ contains the calculation of each index.

**(2) Summary of prediction results of support vector machine classifier**

Run file ‘AUC\_value.m’ and ‘prediction\_index.m’.