**------------------------------------------------------------------------------**

**Readme for the DectICO algorithm package**

**----------------------------------------------------------------------------------------------------------------------**

**============**

**Introduction**

**============**

DectICO is an alignment-free supervised metagenomic sample classification algorithm. It selects the ICO dynamically and classifies the metagenomic samples based on the refined ICO vectors with SVM.

**====================**

**Software Requirement**

**====================**

We require a Windows system with the Matlab and Perl software. Matlab (R2013a 64-bit) and Perl (v5.16.3 64-bit) were used in our experiments.

**=====**

**Usage**

**=====**

DectICO contains mainly three steps:

**Step1:** Extract feature matrix of metagenomic sequencing raw data with the entire sequence feature.

In this step, we provide a perl script to calculate the feature vectors for the metagenomic sequencing raw data. Open the cmd in windows and import:

**>perl integrate\_sample\_feature.pl <fasta.list> <k-mer> <feature type> > <output>**

Input Arguments:

<fasta.list>: The list of names of the metagenomic sequencing files.

<k-mer>: The length of oligonucleotide that used for extracting sequence feature vectors.

<feature type>: The type of sequence feature for characterizing the metagenomes. (1 for the sequence composition and 2 for the ICO)

Output file:

The output file is a feature matrix whose rows represent the feature vector of metagenomic samples. Its name is defined by users.

***Note:*** The list file corresponds to the training samples need to show the samples together which are in different states. For example, we assume the training samples contain 5 positive and 5 negative samples respectively, which are defined as. If the positive sample are , the negative samples have to be  and vice versa.

**Step2:** Select features dynamically and train classifiers with the refined feature matrix.

In this step, users can use “DectICO” function in Matlab to perform both feature selection and training. Open the Matlab console and import:

**>DectICO**

Requirement:

<feature selected ladder file>: A text file named “selection.txt” with the selected feature sizes decreasing from up to bottom.

***Note:*** As shown in our experimental results, the selected feature sets with five components had the best classification performances for all kinds of features on the three metagenomic datasets.  is suggested as the minimum of the size of selected feature set. In addition, our feature selection is a dynamic process, the difference of the sizes of the feature sets between adjacent rounds should not be too large. Therefore we suggest that, for a given kind of sequence feature, the initial entire feature set, with size , is refined about half size recursively each round until the size of selected feature set reaches .

Input Arguments:

<the filename of the training feature maxtrix>: The filename of the training feature matrix extracted in step1.

<the number of positive samples>: The number of positive samples in training set.

<the number of negative samples>: The number of negative samples in training set.

Output file:

The output in this step is a folder named “results” which contains the selected feature set, the LOOCV (leave-one-out cross validation) accuracy and the trained classifier’s model file for each round.

**Step3:** Select the best performed classifier and classify the testing samples.

This step also divides into two processes. First, users need to select the best performed classifier based on the LOOCV accuracies manually. Then we provide a perl script to obtain the refined feature matrix of testing metagenomic samples with the selected feature set file. Open the cmd in windows and import:

**>perl output\_selected\_feature\_matrix.pl <feature matrix> <features selected> <output>**

Input Arguments:

<feature matrix>: The feature matrix of metagenomes with the entire features

<features selected>: The feature set that has the best performance.

<output>: The name of refined feature matrix with the selected features which is defined by users.

Output file:

The output file is a feature matrix with the selected features.

Second, users can classify the testing samples with the refined feature matrix and the best performed classifier’s model file.

Note that, the classification process in DectICO uses the Libsvm software. Therefore the format of input needs to be appropriate. Here, we provide a perl script to perform this process. Open the cmd in windows and import:

**>perl format\_transform\_SVM\_singlefile\_title.pl <testing file> <parameter> > <output>**

Input Arguments:

<testing file>:The feature matrix file that need to transfer format.

<parameter>: Import 1 or -1 optionally (this argument will be acting for the training set).

Output file:

The output file is the feature matrix file with the appropriate format for Libsvm.

Then users can classify the testing samples by Libsvm. Open the cmd in windows and import:

**> svm-predict.exe <transformed file> <model file> <output>**

Input Arguments:

<transformed file>:The transformed feature matrix file.

<model file>: The model file of the best performed classifier.

Output file:

The output file is the result of classification.

**=======**

**Example**

**=======**

Uncompress the “examples” file in the folder which stores the software.

Open cmd and import:

**>perl integrate\_sample\_feature.pl list\_train.txt 8 2 > train.txt**

Open matlab and import:

**>DectICO**

and import following parameters in turn according to the prompts:

Please input the filename of the training feature maxtrix: **train.txt**

Please input the number of positive samples: **5**

Please input the number of negative samples: **5**

Choose the best performed classifier based on the LOOCV accuracy files in the “results” folder and copy the model file of the classifier to the folder which contains the Libsvm software.

Open cmd and import:

**>perl integrate\_sample\_feature.pl list\_test.txt 8 2 > test.txt**

**>perl output\_selected\_feature\_matrix.pl test.txt features\_5.txt** **selected\_test.txt**

**>perl format\_transform\_SVM\_singlefile\_title.pl selected\_test.txt 1 > test**

**>svm-predict.exe test model\_5 res.xls**

The classification result is included in res.xls.

**========**

**Statement**

**========**

The DectICO source builds on the following software:

kernelPLS: <https://github.com/sqsun/kernelPLS>

Libsvm (3.17): <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2011;2(3):27.

Sun, S., Peng, Q. and Shakoor, A. A kernel-based multivariate feature selection method for microarray data classification. *PloS one* 2014;9(7):e102541.