User Guide of DeepDigest

Version 1.7.0

Last revised February 13, 2023

Jinghan Yang

Contents:

1 Introduction

DeepDigest is a Python-based command line tool, which integrates convolutional neural networks (CNNs) and long-short term memory (LSTM) networks to predict the proteotypic cleavage sites for eight commonly used proteases including trypsin, ArgC, chymotrypsin, GluC, LysC, AspN, LysN and LysargiNase. DeepDigest is freely available at http://fugroup.amss.ac.cn/software/DeepDigest/DeepDigest.html.

2 Installation

2.1 Environment Requirements

DeepDigest is a program based on Keras using TensorFlow backend. Here, Python (3.5), TensorFlow (1.10.0) and Keras (2.2.4) are required. Users should also make sure that all the following packages are installed in the Python environment: os, sys, re, getopt, numpy (≥1.14.5). For convenience, we strongly recommend users to install the Anaconda 3 version (64-bit) or above in your local computer, and all the packages can be installed through pip.

2.2 Downloading DeepDigest

DeepDigest can be freely downloaded from http://fugroup.amss.ac.cn/software/DeepDigest/DeepDigest.html. Users can download the release version of DeepDigest "DeepDigest.zip" (Figure 1.a), the help document "User Guide of DeepDigest.pdf" (Figure 1.b) and the test dataset "TestData.zip" (Figure 1.c) from this website.

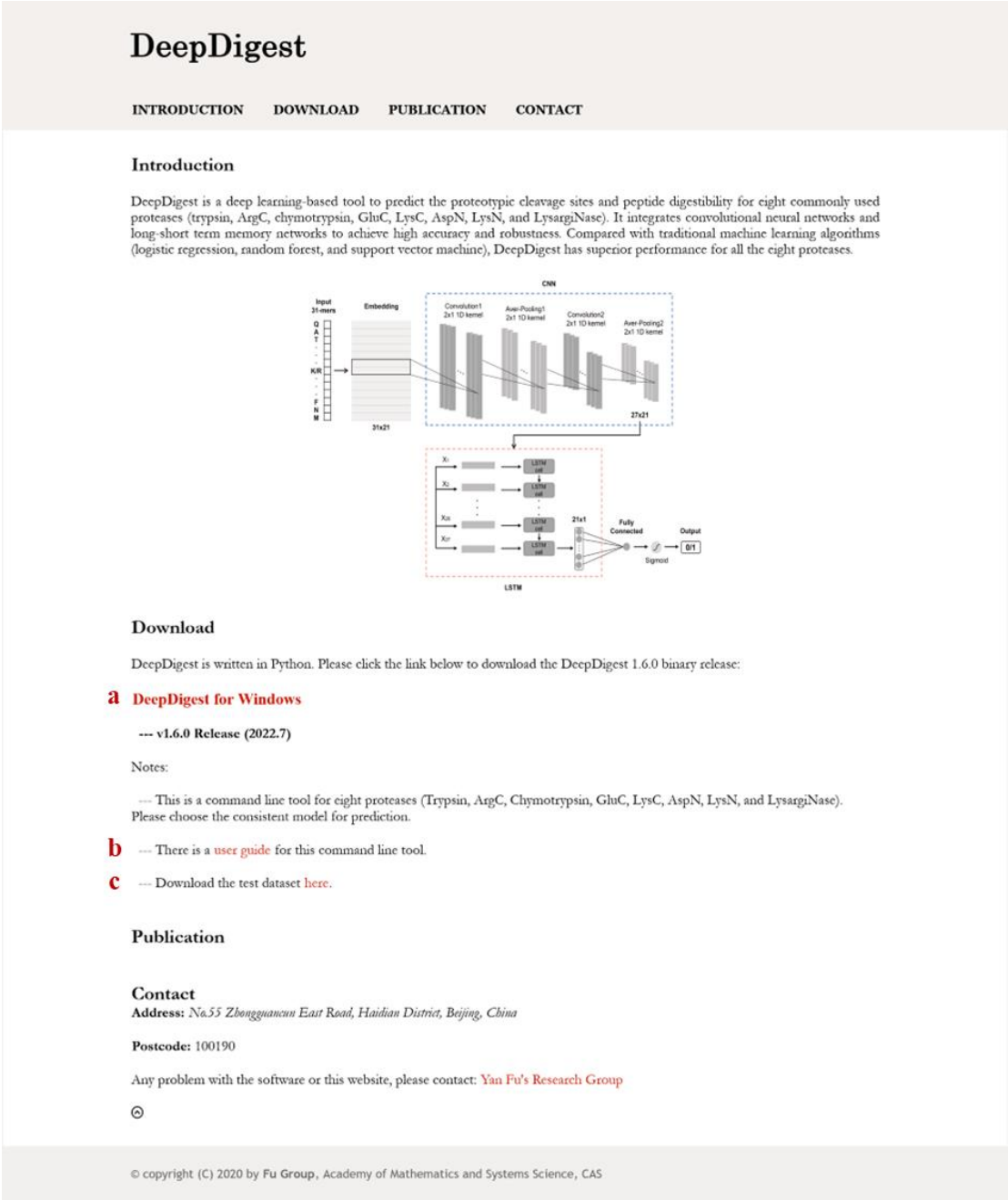

Figure 1. The screenshot of the DeepDigest website.

3 Running DeepDigest

3.1 Parameters

Users can set the custom parameters which are annotated in Table 1:

Table 1. Annotations of the parameters.

Parameter name	Meaning

input	The path of protein sequence file in FASTA format (.fasta)
output	The path of output file in TXT format (.txt)
regular	The regular expression used to extract the protein id (default: ">(.*?)\s")
protease	The digestion protease (default: Trypsin)
missed_cleavages	The maximum number of missed cleavages allowed in each theoretical peptide fragment (default: 2)
min_len	The minimum length of the theoretical peptide fragments (default: 7)
max_len	The maximum length of the theoretical peptide fragments (default: 47)

3.2 Running the command

Open the command interpreter "cmd.exe" and run DeepDigest in the path of the command line tool by the following format:

>python the_main.py --input=*the path of protein sequence file* --output=*the path of output file* --regular=">(.*?)\s" --protease=Trypsin --missed_cleavages=2 --min_len=7 --max_len=47

An example is as follows (Figure 2.):

```
E:\DeepDigest>python the_main.py --input=E:\DeepDigest\nextprot-sparql-entry_PE3.fasta
 --output=E:\DeepDigest\PredictResultsOfPE3_Trypsin.txt --regular="(.*?)\s" --proteas
e=Trypsin --missed_cleavages=2 --min_len =7 --max_len=47
```

Figure 2. Illustration of running DeepDigest.

4 Output file formats

Once the calculation is done, DeepDigest generates a .txt file in the output directory. The detailed description of each column in this file is shown in Table 2.

Table 2. Descriptions of headers in the result file.

Name	Description

Protein id	The identity of the protein from which the peptide is digested
Peptide sequence	The sequence of the theoretical digested peptide
Digestibility of the N-terminal site	The predicted cleavage probability of the cleavage site on the N-terminal of the peptide
Digestibility of the C-terminal site	The predicted cleavage probability of the cleavage site on the C-terminal of the peptide
Digestibility of the missed site(s)	The predicted cleavage probabilities of the missed cleavage sites in the peptide