User guide of gkm-DNN

1. Count the gapped k-mer frequency vector

To count the gapped k-mer frequency vector (gkm-fv), we provide two R scripts. To use these two scripts, you may first need to install three packages namely ‘data.table’, ‘SparseM’ and ‘Stringr’. In the future, we may give an R package which will automatically do these things. Now, we only provide two scripts ‘quick_count.R’ and ‘custom_gkm_example.R’ to count the gkm-fv.

Typical use is: Rscript script.R [args].

The first argument is the needed gkm-fv.RData, which is in the gkmcount folder. The second argument is the name (including path) of the fasta file, the third argument is the name of the output file and the last argument is the number of cores (threads) to use in the script. E.g.

Rscript quick_count.R gkm-fv.RData example.fa example.csv 4

quick_count.R

This script is used to count the gkm-fv (l = 7, k = 5) of DNA sequences. Attention that it will automatically count the results for double-strands.

custom_gkm_example.R

This script is used to count any gkm-fv of DNA sequences (e.g. l = 6, k = 4). Do not run the code in this script. They are just examples to teach you how to use these functions. See source codes for more details.

2. Train a feedforward neural network

Before using gkm-DNN to train a neural network, you should install CUDA 8.0 (gpu calculation) and DL4J. Specifically, just add the dependencies in the POM.xml file into your own POM.xml file of your maven project.

Here, I provide five classes to implement the gkm-DNN. For more information and usage, please see source codes and https://deeplearning4j.org/. The classes are:

MLPBuilder

This class is only used to configure the neural network model as illustrated in the paper.

PresaveData

This is a simple application to transform csv data into binary format. Currently it only support binary classification datasets.

Train

It is a simple example to configure and train the models using early stopping.

PredictDataset

This class is an example to predict a dataset which you know its label (typically for cross validation).

PredictCSV

This class is an example to predict a CSV which you don't know the labels. All the columns in the csv file ought to be input. This is typically used in real application where you don’t know the label of your desired sample.