**Manual for DHS-WE**

1. **INTRODUCTION**

The program is developed for DNase I hypersensitive sites prediction by the genetic algorithm-based weighted ensemble method. Before the prediction, please make sure your computer has installed the python package: repDNA (<http://bioinformatics.hitsz.edu.cn/repDNA/>).

Table 1. 22 sequence-derived features

|  |  |  |
| --- | --- | --- |
| Index | Feature | parameters |
| 1 | 2-kmer | No parameters |
| 2 | 3-kmer | No parameters |
| 3 | 4-kmer | No parameters |
| 4 | 5-kmer | No parameters |
| 5 | 6-kmer | No parameters |
| 6 | Reverse Compliment 2-mer (2-RevcKmer) | No parameters |
| 7 | Reverse Compliment 3-mer (3-RevcKmer) | No parameters |
| 8 | Reverse Compliment 4-mer (4-RevcKmer) | No parameters |
| 9 | Reverse Compliment 5-mer (5-RevcKmer) | No parameters |
| 10 | Reverse Compliment 6-mer (6-RevcKmer) | No parameters |
| 11 | (3, m)-Mismatch Profile | m=1 |
| 12 | (4, m)-Mismatch Profile | m=1 |
| 13 | (5, m)-Mismatch Profile | m=1 |
| 14 | (6, m)-Mismatch Profile | m=1 |
| 15 | (2, w)-Subsequence Profile | w=0.5 |
| 16 | (3, w)-Subsequence Profile | w=0.7 |
| 17 | DACC | lag=1 |
| 18 | Parallel Correlation Pseudo Dinucleotide Composition (PCPseDNC) | λ=2,w=0.2 |
| 19 | Series Correlation Pseudo Dinucleotide Composition (SCPseDNC) | λ=1,w=0.1 |
| 20 | Parallel Correlation Pseudo Trinucleotide Composition (PCPseTNC) | λ=6,w=0.1 |
| 21 | Series Correlation Pseudo Trinucleotide Composition (SCPseTNC) | λ=6,w=0.1 |

1. **DATASETS**

The benchmark dataset is included in “/Dataset/” with the format of “.fasta”. The benchmark dataset contains 280 experimentally verified DHSs sequences and 737non-DHSs sequences.

1. **CODES**

For the individual feature-based models and the genetic algorithm-based weighted ensemble model, you need to first encode the sequences to various features, and then use the models for DHSs prediction. The predicted results are saved in the file of “.txt”.

The code is written in MATLAB. To predict the DHSs, please type the following command in MATLAB environment:

issuccess=DHS\_predict(inputfile,outputfile),

pleat note that the inputfile is fasta format.

For example:

issuccess=DHS\_predict(‘test.fasta’,’test\_out.txt’)