PseDNA-Pro

General Claimed:

(A) proposed a feature vector composed of three kinds of sequence-based features, including overall amino acid composition, pseudo amino acid composition (PseAAC) proposed by Chou and physicochemical distance transformation.

(B) The proteins in the training set and test set were transformed into fixed-dimension feature vectors. The feature vectors were fed into Support Vector Machine (SVM) for DNA-binding protein identification.

(C) trained a prediction model using SVM with linear kernel. Used Lib-SVM package.The radial basis function (RBF) is taken as the kernel functions. The two parameters C and t were optimized on the benchmark dataset by adopting the grid tool provide by LIBSVM.

(D) achieves accuracy values:

jackknife test: 80.05% (for both λ=1 and λ=2)

jackknife test: 76.55% (extended benchmark dataset)

independent test: PseDNA-Pro (λ=1) : 82.22% and PseDNA-Pro (λ=2) : 83.33%

Features Used:

1. Overall Amino Acid Composition (OAAC): here OAAC is defined as the occurrence frequencies of 20 standard amino acids.

2. Pseudo Amino Acid Composition (PseAAC): considered seven physicochemical properties, including the Transfer energy, Hydrophobicity value, Packing density, Nature of the accessible and buried surfaces, Shape and surface features, Alpha-helix indices, and Helix-coil equilibrium constant.

3.Physicochemical Distance Transformation (PDT): considers all the 531 meaningful amino acid indices extracted from AAIndex1 database. A protein sequence S with L amino acids can be represented as a 531 dimensional vector

Performance Measures:

(A) Jackknife on the benchmark dataset for both PseDNA-Pro (λ=1) and PseDNA-Pro (λ=2): Accuracy (80.05%), Sensitivity (63.43%), Specificity (89.06%), MCC (0.55)

(B) Independent dataset:

PseDNA-Pro (λ=1) : Acc (82.22%), SE (76.54%), SP (86.86%), MCC (0.63).

PseDNA-Pro (λ=2) : Acc (83.33%), SE (76.54%), SP (88.88%), MCC (0.66).

(C) Extended benchmark dataset through Jackknife validation:

(The extended benchmark dataset contains 525 DNA-binding proteins and 550 non-DNA-binding proteins.)

PseDNA-Pro : ACC (76.55%), MCC (0.53), SE (79.61%),SP (73.63%)