DPP-PseAAC: A DNA-binding protein prediction model using Chou's general PseAAC

BibTeX:

```
@article{rahman2018dpp,
  title={DPP-PseAAC: A DNA-binding protein prediction model using Chou's general PseAAC},
  author={Rahman, M Saifur and Shatabda, Swakkhar and Saha, Sanjay and Kaykobad, M and Rahman, M Sohel},
  journal={Journal of theoretical biology},
  volume={452},
  pages={22--34},
  year={2018},
  publisher={Elsevier}
}
```

General Claimed:

(A) model extracts meaningful information directly from the protein sequences, without any dependence on functional domain or structural information.

(B) employed RF model to rank the features. Then, used SVM-RFE (Recursive Feature Elimination) method to extract an optimal set of features; Top 289 features selected from extracted features.

(C) trained a prediction model using SVM with linear kernel.

(D) achieves accuracy values:

10-fold cross-validation test: 93.21%,
jackknife test: 95.91% and
independent test: 77.42%.

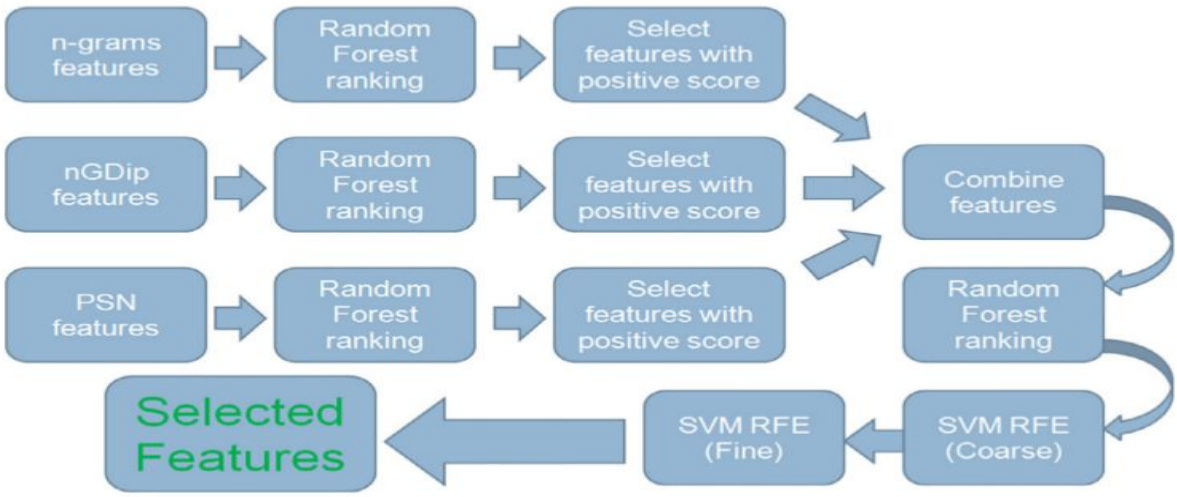


Fig. 2. Steps in feature selection.

Features Used:
1. Amino Acid Composition (AAC)
2. Di-peptides: AA, AC, AD, AE ...
3. Tri-peptides: AAA, AAC, AAD, AAE ...
4. n-gapped-dipeptides: A_A, A__A, A___A, A____A ... (upto 25 gaps!) and
5. Position specific n-grams (n=1,2,3)

Total extracted features: 29,679

Performance Measures:
(A) Jackknife cross-validation on the PDB1075 dataset:
Accuracy (95.91%), Sensitivity (94.10%), Specificity (97.64%), MCC (0.92), auROC (0.9884).

(B) 10-fold cross validation on the PDB1075 dataset:
Accuracy (93.21%), sensitivity (87.81%), specificity (98.36%), MCC (0.87), auROC (0.98).

(C) Jackknife cross validation on the PDB186 independent-dataset:
Accuracy (77.42%), sensitivity (83.87%), specificity (70.97%), MCC (0.553), auROC (0.7986).

Tools: http://77.68.43.135:8080/DPP-PseAAC
Codes: https://github.com/srautonu/DNABinding

StackDPPred: A Stacking based Prediction of DNA-binding Protein from Sequence

BibTeX:
```
@article{mishra2018stackdppred,
  title={StackDPPred: A Stacking based Prediction of DNA-binding Protein from Sequence},
  author={Mishra, Avdesh and Pokhrel, Pujan and Hoque, Md Tamjidul and Hancock, John},
  journal={Bioinformatics},
  volume={1},
  pages={9},
  year={2018}
}
```

General Claimed:
(A) The classifiers of the first-stage (base-classifier) and second-stage (meta-classifier).
A pool of base-classifiers are employed in the first-stage. Then, using meta-classifier in the second-stage, the outputs of the base-classifiers are combined with the aim of reducing the generalization error.

(B) To select the best combination of classifiers to be used as the base-classifiers, we first analyze the performance of six different machine learning methods, SVM, LogReg, KNN, RDF, BAG and ET, on the benchmark dataset.

(B) train the ensemble of predictors at the first-stage Then, they combine the output of the predictors at the base-layer using another SVM at
the second-stage. As a result, the meta-learner SVM of the StackDPPred achieves an ACC of 89.96%, MCC of 0.7990 and AUC of
0.9449 on a benchmark dataset, whereas the base-learner SVM achieves an ACC of 80.43% and MCC of 0.60849.

(C) This achievement, allows us to conclude that stacking technique helps reduce the generalization error
and thus can improve accuracy significantly.

(D) Stacking Method: includes SVM, Logistic Regression, KNN and RF;

(E) DNA-binding proteins using machine learning, involves two important steps:
(i) extraction of relevant features; and
(ii) selection of an appropriate classification algorithm.

Features Used:
1. Position Specific Scoring Matrix (PSSM) and
2. Reside Wise Contact Energy Matrix (RCEM)

Performance Measures:
(A) achieves results on PDB1075 using jackknife-test:
ACC 89.96%, MCC 0.799, and AUC 94.50%.

(B) achieves results on PDB186 using jackknife-test:
ACC (86.55%) Sensitivity (92.47%) Specificity (80.64%) MCC (0.7363) AUC (0.8878).

DBPPred: Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naı¨ve Bayes

BibTeX:

@article{lou2014sequence,
  title={Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes},
  author={Lou, Wangchao and Wang, Xiaoqing and Chen, Fan and Chen, Yixiao and Jiang, Bo and Zhang, Hua},
  journal={PLoS One},
  volume={9},
  number={1},
  pages={e86703},
  year={2014},
  publisher={Public Library of Science}
}

General Claimed:
(A) performing the feature rank using random forest (to rank the features using Gini importance) and the wrapper-based feature selection using forward best-first search strategy
(B) RF filter GNB Wrapper (56 features)
(C) used GNB as a classifier
(D) The features comprise information from the primary sequence, the predicted secondary structure, the predicted relative solvent accessibility, and the position specific scoring matrix.

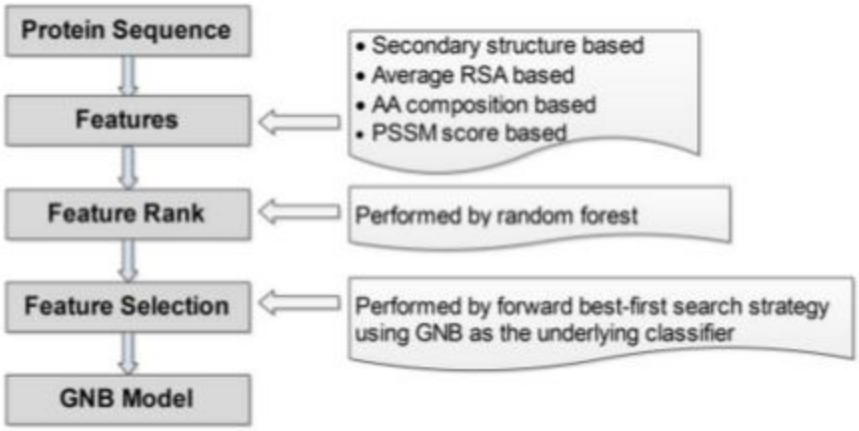


Figure 1. The flowchart of the proposed method.
doi:10.1371/journal.pone.0086703.g001

Features Used:
AAC, PredSS, PredRSA Auto-correlation coefficients of PSSM, Percentile values of PSSM scores ;

---

iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition

BibTeX:

```
@article{liu2014idna,
  title={iDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition},
  author={Liu, Bin and Xu, Jinghao and Lan, Xun and Xu, Ruifeng and Zhou, Jiyun and Wang, Xiaolong and Chou, Kuo-Chen},
  journal={PloS one},
  volume={9},
  number={9},
  pages={e106691},
  year={2014},
  publisher={Public Library of Science}
}
```

General Claimed:
(A) To avoid dimension disaster and reduce computational time, the reduced amino acid alphabet strategy was adopted.
(B) That is why the new predictor can outperform the existing predictors in identifying DNA-binding proteins with less computational time.
(C) It is anticipated that the iDNA-Prot|dis predictor will become a high throughput tool for both basic research and drug development.
(C) 602 features selected from 1220;
(D) used SVM (RBF) classifier ;

Features Used:
(A) Amino acid distance-pair coupling Amino acid reduced alphabet profile

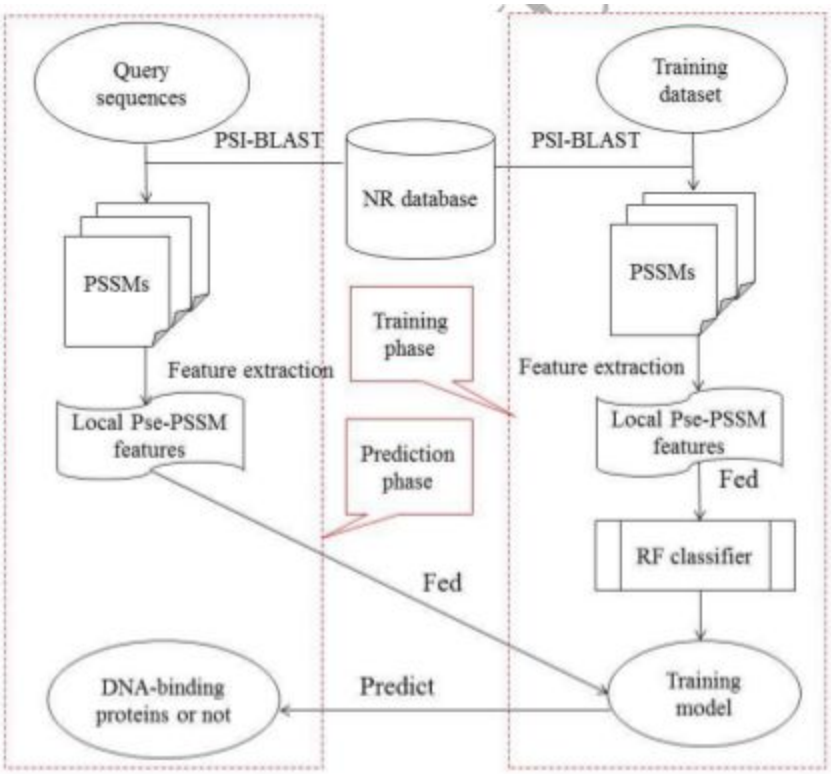
---

Local-DPP: An Improved DNA-binding Protein Prediction Method by Exploring Local Evolutionary Information

BibTeX:

@article{wei2017local,
  title={Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information},
  author={Wei, Leyi and Tang, Jijun and Zou, Quan},
  journal={Information Sciences},
  volume={384},
  pages={135--144},
  year={2017},
  publisher={Elsevier}
}

General Claimed:
(A) proposes a novel feature representation algorithm that efficiently extracts the local features from the profiles (PSSM) ;
(A) combines the local Pse-PSSM features with the random forest classifier ;
(C) features derived from local regions are more discriminative than features derived from the whole region ;
(C) use RF classifier;
(D) 120 features selected by gini index;



Features Used:
(A) Local Pse-PSSM

Kmer1+ACC: Identification of DNA-binding proteins by auto-cross covariance transformation

BibTeX:
@inproceedings{dong2015identification,
  title={Identification of DNA-binding proteins by auto-cross covariance transformation},
  author={Dong, Qiwen and Wang, Shanyi and Wang, Kai and Liu, Xuan and Liu, Bin},
  booktitle={Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on},
  pages={470--475},
  year={2015},
  organization={IEEE}
}

General Claimed:
(A) novel framework is proposed to encode the protein sequences with different lengths into fixed-length vector by using Auto-Cross Covariance transformation (ACC) ;
(B) combines the support vector machine and the auto-cross covariance transformation ;
(C) used SVM classifier ;

Features Used:
(A) ACC, kmer composition, Physico-chemical properties ;

DNA-Prot: Identification of DNA Binding Proteins from Protein Sequence Information using Random Forest

BibTeX:
@article{kumar2009dna,
  title={DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest},
  author={Kumar, K Krishna and Pugalenthi, Ganesan and Suganthan, Ponnuthurai N},

journal={Journal of Biomolecular Structure and Dynamics},
volume={26},
number={6},
pages={679--686},
year={2009},
publisher={Taylor \& Francis}
}

General Claimed:
(A) Best-first-search explores the space of attribute subsets by using the greedy hill-climbing augmented with the backtracking ;
(B) correlation-based feature subset selection method (CFSS) for the feature selection ;
(C) used RF classifier ;
(D) encoded by 116 features, and selected 20 among them ;

Features Used:
(A) Frequency of amino acid/amino acid groups, hydrophobic, hydrophlic, neutral residues, PredSS from PSIPRED, Amino acid physico-chemical properties, Split sliding 10 residue windows.

Datasets: In the benchmark dataset containing 823 DNA-binding proteins and 823 non DNA-binding proteins ;

---

iDNAPro-PseAAC: DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation

BibTeX:
@article{liu2015dna,
title={DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation},
author={Liu, Bin and Wang, Shanyi and Wang, Xiaolong},
journal={Scientific reports},
volume={5},
pages={15479},
year={2015},
publisher={Nature Publishing Group}
}

General Claimed:
(A) able to incorporate the global or long range sequence-order effects
(B) use SVM (with RBF kernel) classifier ;
(C) The evolutionary information imbedded in the profile-based protein representation is employed by iDNAPro-PseAAC
(C) Profile-based protein representation PseAAC (λ=3) ;
(D) 23 features selected ;

---

iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model

BibTeX:

```
@article{lin2011idna,
  title={iDNA-Prot: identification of DNA binding proteins using random forest with grey model},
  author={Lin, Wei-Zhong and Fang, Jian-An and Xiao, Xuan and Chou, Kuo-Chen},
  journal={PloS one},
  volume={6},
  number={9},
  pages={e24756},
  year={2011},
  publisher={Public Library of Science}
}
```

General Claimed:
(A) use RF classifier
(B) 23 features

DNAbinder: Identification of DNA-binding proteins using support vector machines and evolutionary profiles

BibTeX:

@article{kumar2007identification,
  title={Identification of DNA-binding proteins using support vector machines and evolutionary profiles},
  author={Kumar, Manish and Gromiha, Michael M and Raghava, Gajendra PS},
  journal={BMC bioinformatics},
  volume={8},
  number={1},
  pages={463},
  year={2007},
  publisher={BioMed Central}
}

General Claimed:
(A) It has been observed that PSSM based models perform better than any other models by 3–7% on all the datasets including independent and
realistic datasets.
(B) 400 features used
(C) used SVM classifier

HMMBinder: DNA-Binding Protein Prediction Using HMM Profile Based Features

BibTeX:
@article{zaman2017hmmbinder,
   title={HMMBinder: DNA-Binding Protein Prediction Using HMM Profile Based Features},
   author={Zaman, Rianon and Chowdhury, Shahana Yasmin and Rashid, Mahmood A and Sharma, Alok and Dehzangi, Abdollah and Shatabda, Swakkhar},
   journal={BioMed research international},
   volume={2017},
   year={2017},
   publisher={Hindawi}
}

General Claimed:
(A)

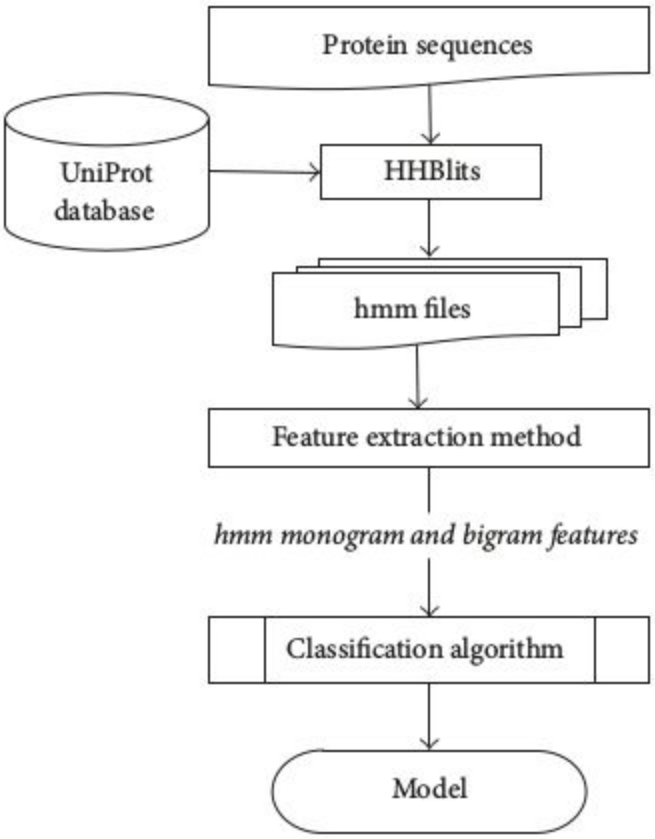


FIGURE 1: System diagram of HMMBinder.

Feature Used:
(A) HMM profile based features of a protein sequence.