Journal of Theoretical Biology 452 (2018) 22–34

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

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

M. Saifur Rahman[a], Swakkhar Shatabda[b], Sanjay Saha[c], M. Kaykobad[a], M. Sohel Rahman[a,*]

[a] Department of CSE, BUET, ECE Building, West Palasi, Dhaka 1205, Bangladesh
[b] Department of Computer Science and Engineering, United International University, House 80, Road 8A, Dhanmondi, Dhaka 1209, Bangladesh
[c] Department of Computer Science and Engineering, University of Asia Pacific, 74/A Green Road, Dhaka 1215, Bangladesh

ARTICLE INFO

Article history:
Received 17 March 2018
Revised 21 April 2018
Accepted 4 May 2018

Keywords:
DNA binding
Classification
Prediction
Support Vector Machine
Random Forest
PseAAC

ABSTRACT

A DNA-binding protein (DNA-BP) is a protein that can bind and interact with a DNA. Identification of DNA-BPs using experimental methods is expensive as well as time consuming. As such, fast and accurate computational methods are sought for predicting whether a protein can bind with a DNA or not. In this paper, we focus on building a new computational model to identify DNA-BPs in an efficient and accurate way. Our model extracts meaningful information directly from the protein sequences, without any dependence on functional domain or structural information. After feature extraction, we have employed Random Forest (RF) model to rank the features. Afterwards, we have used Recursive Feature Elimination (RFE) method to extract an optimal set of features and trained a prediction model using Support Vector Machine (SVM) with linear kernel. Our proposed method, named as *DNA-binding Protein Prediction model using Chou's general PseAAC (DPP-PseAAC)*, demonstrates superior performance compared to the state-of-the-art predictors on standard benchmark dataset. DPP-PseAAC achieves accuracy values of 93.21%, 95.91% and 77.42% for 10-fold cross-validation test, jackknife test and independent test respectively. The source code of DPP-PseAAC, along with relevant dataset and detailed experimental results, can be found at https://github.com/srautonu/DNABinding. A publicly accessible web interface has also been established at: http://77.68.43.135:8080/DPP-PseAAC/.

1. Introduction

A DNA-binding protein (DNA-BP) is a protein that can bind and interact with a DNA. Such a protein is composed of DNA binding domains that include transcription factors, nucleases and histones. The transcription factors modulate the process of transcription, while the nucleases can cleave DNA molecules. Histones, on the other hand, are involved in chromosome packaging in the cell nuclei. Fig. 1 shows examples of protein DNA binding interactions: in the left figure, a transcription factor is bound to a DNA, while in the right figure, the restriction enzyme EcoRV is interacting with its target DNA.

The DNA-BPs thus perform two main functions: firstly, they organize and compact the DNA and secondly, they regulate and affect various cellular processes like transcription, DNA replication, recombination, repair and modification. Therefore, the DNA-BPs can potentially be used for drug development in treating genetic diseases and cancers Gurova (2009) and Leung et al. (2013). This is why developing efficient and highly accurate methods to identify DNA-BPs is a very important research problem in the field of molecular biology. Traditionally, the DNA-BPs have been identified through different experimental methods. These include filter binding assays Helwa and Hoheisel (2010), genetic analysis Freeman et al. (1995), X-ray crystallography Chou et al. (2003), chromatin immunoprecipitation on microarrays Buck and Lieb (2004) etc. However, as these experimental methods are costly and time consuming, researchers have started to rely on computational methods to identify DNA-BPs. These methods can largely be categorized into two groups: structure based methods and sequence based methods.

Structure-based methods depend on the structural information of the protein sequences. These include high-resolution 3D structure, accessible surface area, torsion angles, structure motifs etc. Stawiski et al. (2003) did the pioneering work in identifying DNA-BPs using structural information. They extracted features from the detailed atomic structure of the protein and then

* Corresponding author.
E-mail addresses: mrahman@cse.buet.ac.bd (M.S. Rahman), swakkhar@cse.uiu.ac.bd (S. Shatabda), sanjay@uap-bd.edu (S. Saha), kaykobad@cse.buet.ac.bd (M. Kaykobad), msrahman@cse.buet.ac.bd (M.S. Rahman).

https://doi.org/10.1016/j.jtbi.2018.05.006
0022-5193/

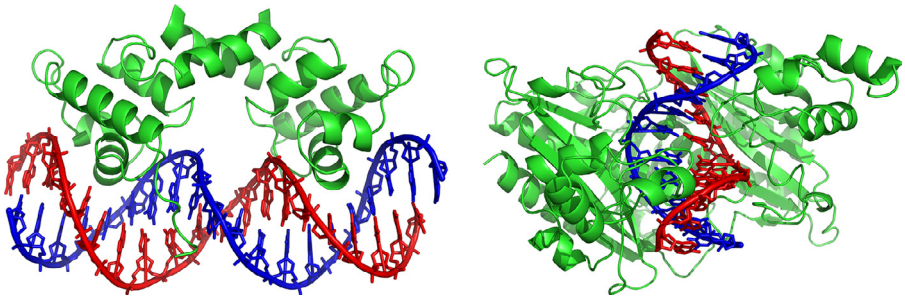

Fig. 1. DNA-binding proteins bound to respective target DNAs. **(Left)** The lambda repressor helix-turn-helix transcription factor bound to its DNA target. Created from PDB 1LMB. Image source: Zephyris (2018b). **(Right)** The restriction enzyme EcoRV in a complex with its substrate DNA. Created from PDB 1RVA. Image source: Zephyris (2018a).

employed a three-layer artificial neural network (ANN). Ahmad and Sarai (2004), on the other hand, used a two-layer neural network with features calculated solely from bulk electrostatic properties. Szilágyi and Skolnick (2006) subsequently proposed a fast and efficient method to predict DNA-BPs from only the amino acid sequences and low-resolution, C^{α}-only protein models. Their predictor is available as a web-server called *DNABIND*. Gao and Skolnick (2008) proposed *DBD-Hunter* that applies structural alignment and evaluation of a statistical potential to identify DNA-BPs. Gao and Skolnick (2009) subsequently proposed *DBD-Threader*, for the prediction of DNA-binding domains and associated DNA-binding protein residues. While this method uses a template library composed of DNA-protein complex structures, it requires only the target protein's sequence for its classification. This independence from structural information makes the predictor very useful, while its performance remains comparable with DBD-Hunter. Examples of other structure-based methods can be found in Zhao et al. (2010), Nimrod et al. (2010), Zhou and Yan (2011) and Szabóová et al. (2012).

Structure-based predictors are applicable only when the structural information of a candidate protein is known. While the post-genomic era witnesses a rapid growth in sequence known proteins, the structure of many of these proteins still remain undiscovered. The predictors that solely rely on structural information of proteins are thus limited in their use. Sequence based methods, on the other hand, attempt to identify the DNA-BPs from the amino acid sequence by extracting various discriminating features. Some predictors may additionally rely on some structural features for improved prediction accuracy when the protein structure is known. Examples of prominent sequence based predictors of DNA-BPs can be found in Kumar et al. (2007), Fang et al. (2008), Kumar et al. (2009), Nanni and Lumini (2009), Shao et al. (2009), Lin et al. (2011), Zhao et al. (2012), Zou et al. (2013), Lou et al. (2014), Xu et al. (2014), Song et al. (2014), Liu et al. (2015c), Dong et al. (2015), Liu et al. (2015d), Xu et al. (2015), Motion et al. (2015), Im et al. (2015), Waris et al. (2016), Zhou et al. (2016), Paz et al. (2016), Wei et al. (2017) and Chowdhury et al. (2017).

Kumar et al. (2007) used evolutionary information from the Position Specific Scoring Matrix (PSSM) for protein representation. The PSSM profile of each protein was generated from PSI-BLAST Altschul et al. (1997) by searching the non-redundant (nr) protein database using three iterations with e-value cutoff set to 0.001. They applied Support Vector Machine (SVM) Boser et al. (1992) as the learner. Available as a webtool called *DNAbinder*, the performance of their predictor depends on the quality of PSSM profiles, which is heavily dependent on the database being searched for homology information. To eliminate this dependency, *DNA-Prot* was proposed by another group Kumar et al. (2009). This predictor used features such as frequency of amino acid residues and groups, predicted secondary structure (PredSS) information from PSIPRED McGuffin et al. (2000), physico-chemical properties from AAIndex database Kawashima et al. (2007). To reduce the feature vector size, they applied Correlation-based feature subset selection method (CFSS).

Lin et al. (2011) incorporated the Grey model Julong (1989) parameters in the general form of Chou's PseAAC Chou (2011) for protein sequence representation. They then trained their model, *iDNA-Prot*, using Random Forest (RF) Breiman (2001). Lou et al. (2014) introduced a predictor called *DBPPred*, where amino acid composition, PSSM scores, PredSS and predicted relative solvent accessibility (PredRSA) were used as features. They then used Random Forest to rank the features, followed by a wrapper method. They used Gaussian Naïve Bayes (GNB) as the final classifier. They compared their predictor with prior ones using an independent dataset called *PDB186*, comprising equal number of DNA-binding and non DNA-binding proteins. This dataset has subsequently been used in performance evaluation of many other predictors.

Liu et al. (2014) used amino acid distance-pair coupling information into Chou's general form of PseAAC Chou (2011). To reduce the dimension of the feature vector and to speed up the prediction process, they also used amino acid reduced alphabet profile Peterson et al. (2009). They then applied SVM with RBF kernel to produce the prediction tool called *iDNA-Prot|dis*. To train and assess their predictor using cross-validation, they prepared a stringent balanced dataset of 1075 protein samples. This benchmark dataset has subsequently been referred to as *PDB1075* and has been widely used in literature for cross-validation. We have also used this dataset in our work and provide a detailed description of the dataset later in the paper. In addition to preparation of the benchmark dataset, a key contribution of Liu et al.'s work was re-implementation of major earlier predictors and measuring their cross-validation performance using this benchmark dataset. This paved the way for subsequent predictors to be compared with prior art in an apple for apple comparison.

In 2015, Liu et al. (2015c) presented another predictor called *iDNAPro-PseAAC*. They used profile-based representation of the protein sequence and then used PseAAC with the 3rd order sequence-order effect. Dong et al. (2015) used Auto-Cross Covariance (ACC) transformation with amino acid k-mer compositions and physicochemical properties. They then used SVM to train the predictor, widely known as *Kmer1 + ACC*. Wei et al. proposed *Local-DPP* Wei et al. (2017), where local pseudo position specific scoring matrix (Local Pse-PSSM) features have been used. In this approach, the locally conserved protein information is captured by fragmenting the PSSMs into several equally sized sub-PSSMs. Finally, all the local features are fed into the Random Forest algorithm to learn the classification model.

Very recently, Chowdhury et al. developed *iDNAProt-ES* Chowdhury et al. (2017), that utilizes both the evolutionary