entropy

MDPI

Article

A Novel Sequence-Based Feature for the Identification of DNA-Binding Sites in Proteins Using Jensen–Shannon Divergence

Truong Khanh Linh Dang [1], Cornelia Meckbach [2], Rebecca Tacke [2], Stephan Waack [1] and Mehmet Gültas [1,*]

[1] Institute of Computer Science, University of Göttingen, Göttingen 37077, Germany; ldang1@informatik.uni-goettingen.de (T.K.L.D.); waack@informatik.uni-goettingen.de (S.W.)

[2] Institute of Bioinformatics, University Medical Center Göttingen, Göttingen 37077, Germany; c.meckbach@bioinf.med.uni-goettingen.de (C.M.); rebecca.tacke@stud.uni-goettingen.de (R.T.)

* Correspondence: gueltas@informatik.uni-goettingen.de; Tel.: +49-551-39-172055

Academic Editors: Carlos M. Travieso-González and Jesús B. Alonso-Hernández
Received: 30 July 2016; Accepted: 20 October 2016; Published: 24 October 2016

Abstract: The knowledge of protein-DNA interactions is essential to fully understand the molecular activities of life. Many research groups have developed various tools which are either structure- or sequence-based approaches to predict the DNA-binding residues in proteins. The structure-based methods usually achieve good results, but require the knowledge of the 3D structure of protein; while sequence-based methods can be applied to high-throughput of proteins, but require good features. In this study, we present a new information theoretic feature derived from Jensen–Shannon Divergence (JSD) between amino acid distribution of a site and the background distribution of non-binding sites. Our new feature indicates the difference of a certain site from a non-binding site, thus it is informative for detecting binding sites in proteins. We conduct the study with a five-fold cross validation of 263 proteins utilizing the Random Forest classifier. We evaluate the functionality of our new features by combining them with other popular existing features such as position-specific scoring matrix (PSSM), orthogonal binary vector (OBV), and secondary structure (SS). We notice that by adding our features, we can significantly boost the performance of Random Forest classifier, with a clear increment of sensitivity and Matthews correlation coefficient (MCC).

Keywords: entropy; Jensen–Shannon divergence; Random Forest; DNA-binding sites

1. Introduction

Interactions between proteins and DNA play essential roles for controlling of several biological processes such as transcription, translation, DNA replication, and gene regulation [1–3]. An important step to understand the underlying molecular mechanisms of these interactions is the identification of DNA-binding residues in proteins. These residues can provide a great insight into the protein function which leads to gene expression and could also facilitate the generation of new drugs [4,5].

Until now, several groups have published different studies based on either experimental or computational identification of DNA-binding proteins [1,6–11] as well as residues in these proteins [12–23]. However, the usage of experimental approaches for the determination of binding sites is still challenging since they are often demanding, relatively expensive, and time-consuming. To overcome the difficulty of experimental approaches, it is highly desired to develop fast and reliable computational methods for the prediction of DNA-binding residues. For this purpose, several state-of-the-art prediction methods have been developed for the automated identification of those residues. Such methods can be assigned into two main categories: (i) based on the information observed from structure and sequence in a collective manner; (ii) based on the features derived directly

from the amino acid sequence alone (for more detail see reviews [24] and [25]). Although the first type of approaches provides promising information about DNA-binding residues in proteins, their application is difficult due to the limited number of experimentally determined protein structures. In contrast to structure-based approaches, sequence-based methods have been developed by extracting different sequence information features, like amino acid frequency, position-specific scoring matrix (PSSM), BLOSUM62 matrix, sequence conservation, etc. [3,4,18,19,26,27]. Using these features, several machine learning techniques have been applied to construct the classifiers for the prediction of binding residues in proteins. To this end, a variety of support vector machine (SVM) classifiers have been developed in recent studies [2,17–19,23,26,28]. For example, Westhof et al. have recently used an SVM classifier approach in their study, named RBscore (http://ahsoka.u-strasbg.fr/rbscore/), by using the physicochemical and evolutionary features that are linearly combined with a residue neighboring network [2]. Further, SVM algorithms were also applied for the models proposed in BindN [18], DISIS [19], BindN+ [23], DP-Bind [27] using different sequence information features including the biochemical property of amino acids, sequence conservation, evolutionary information in terms of PSSM, the side chain pKa value, hydrophobicity index, molecular mass and BLOSUM62 matrix. In addition, other machine learning classifiers such as neural network models [13,15], naive Bayes classifier [26], Random Forest classifiers (RF) [4,29,30] have been developed based on the features derived from protein sequences. For example, Wong et al. [29] have recently developed a successful method using RF classifier with both DNA and protein derived features to predict the specific residue-nucleotide interactions for different DNA-binding domain families.

Despite the rich literature on the sequence-based methods as mentioned above, to date there is still a need to find suitable feature extraction approaches that can enhance the characteristics of DNA-binding residues and thus help to improve the performance of existing methods for identification of DNA-binding residues in proteins. For this aim, we introduce and evaluate a new information theory-based method for the prediction of these residues using Jensen–Shannon divergence ($\mathbb{JSD}$). As a divergence measure based on the Shannon entropy, $\mathbb{JSD}$ is a symmetrized and smoothed version of the Kullback-Leibler divergence and is often used for different problems in the field of bioinformatics [31–35]. In this study, following the line of Capra et al. [34] we first quantify the divergence between the observed amino acid distribution of a site in a protein and the background distribution of non-binding sites by using $\mathbb{JSD}$. After that, in analogy to our previous studies QCMF [32] and CMF [36], we incorporate biochemical signals of binding residues in the calculation of $\mathbb{JSD}$ that results in the intensification of the DNA-binding residue signals from the non-binding signals.

To demonstrate the performance and functionality of our proposed approach, we apply Random Forest (RF) classifier using our new $\mathbb{JSD}$ based features together with three widely used machine learning features, namely position-specific scoring matrix (PSSM), secondary structure (SS) information, and orthogonal binary vector (OBV) information (see review [24]). Our results show that using $\mathbb{JSD}$ based features, RF classifier reaches an improved performance in identifying DNA-binding residues with a significantly higher Matthews correlation coefficient (MCC) value in comparison to using previous features alone. Although we only applied RF classifier in this study, both of our sequence-based features could be used in other classifiers such as SVM, neural networks, or decision trees.

2. Results

In this study, we introduce new sequence-based features using $\mathbb{JSD}$ to improve the performance of previous machine learning approaches in identification of DNA-binding residues in proteins. For this purpose, we propose new sequence-based features ($f_{\mathbb{JSD}}$ and $f_{\mathbb{JSD}\text{-t}}$) using $\mathbb{JSD}$ in two different ways. First, using $\mathbb{JSD}$, we calculate the divergences between observed amino acid distributions in multiple sequence alignments (MSAs) of proteins under study and the background distribution which is calculated according to amino acid counts at non-binding residue positions in MSAs. In the second step, we transform the observed amino acid distributions with a doubly stochastic matrix (DSM) to

enhance the weak signal of binding sites in proteins which could not be predicted in the first step. Finally, we calculate for each residue in proteins $\mathbb{JSD}$-based scores and use them for the improvement of the performance of machine learning approaches.

To evaluate our new features, we use two frequently considered cut-off distances of 3.5 Å and 5 Å and thus define a residue in a protein as DNA-binding if the distance between at least one atom on its backbone or side chain and the DNA molecule is smaller than the considered cut-off.

The Results section of this study comprises of two parts. First, we investigate the functionality of our new features combining them in Random Forest (RF) classifier with three previous features. The RF classifier is constructed from 4298 positive and 44,805 negative instances extracted from 263 proteins. The performance of the classifier is evaluated using a five-fold cross validation procedure in which we randomly divided the samples into five parts. The assessment is performed by choosing each of these parts as a test set and the remaining four parts as a training set for model selection. Second, to illustrate the usefulness of our new approach for the prediction of DNA-binding residues, we analyzed the proto-oncogenic transcription factor MYC-MAX (PDB-ID: 1NKP) which is a heterodimer protein complex of two proteins. It is important to note that this protein complex is not included in the training dataset.

2.1. Random Forest Classifier

To apply the Random Forest (RF) classifier, we combine our new features ($f_{\mathbb{JSD}}$ and $f_{\mathbb{JSD}\text{-t}}$) with the features f_{PSSM}, f_{OBV}, and f_{SS} which are widely used for the prediction of DNA-binding residues. Our results show that using our features RF classifier reaches an improved performance in identifying DNA-binding sites with clearly higher statistical values (see Tables 1 and 2). Moreover, we individually evaluated the combination of our features with existing features. The results suggest that the classifier with $f_{\mathbb{JSD}\text{-t}}$ feature has provided better sensitivity and comparable Matthews correlation coefficient (MCC) values in comparison to $f_{\mathbb{JSD}}$ feature. However, its specificity is moderately decreased. A further comparison reveals that the usage of our both features together with other features does not affect the performance of the classifier. The details are presented for 3.5 Å in Table 1 and for 5 Å in Table 2 and in Appendix A with the standard error of each of the performance measures over the values obtained in the five iterations (see Tables A1 and A2).

Table 1. Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å. The prediction system was evaluated by five-fold cross validation.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f_{PSSM}	0.292	0.963	0.307	0.777	0.313
f_{PSSM} + $f_{\mathbb{JSD}}$	0.385	0.949	0.349	0.795	0.369
f_{PSSM} + $f_{\mathbb{JSD}\text{-t}}$	0.41	0.939	0.35	0.802	0.377
f_{PSSM} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	0.414	0.94	0.348	0.800	0.376
f_{PSSM} + f_{SS}	0.339	0.958	0.334	0.794	0.338
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$	0.416	0.95	0.378	0.808	0.390
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	0.441	0.94	0.372	0.817	0.401
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	0.439	0.94	0.37	0.814	0.399
f_{PSSM} + f_{OBV} + f_{SS}	0.367	0.968	0.398	0.838	0.413
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$	0.422	0.958	0.409	0.837	0.425
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	0.447	0.95	0.403	0.841	0.431
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	0.444	0.947	0.393	0.835	0.423

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

Table 2. Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å. The prediction system was evaluated by five-fold cross validation.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f_{PSSM}	0.286	0.966	0.350	0.778	0.425
$f_{PSSM} + f_{\mathbb{JSD}}$	0.395	0.95	0.407	0.801	0.487
$f_{PSSM} + f_{\mathbb{JSD}\text{-t}}$	0.418	0.943	0.411	0.807	0.494
$f_{PSSM} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.426	0.942	0.414	0.807	0.497
$f_{PSSM} + f_{SS}$	0.334	0.963	0.386	0.796	0.455
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}}$	0.424	0.951	0.436	0.814	0.513
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.448	0.944	0.438	0.820	0.520
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.445	0.944	0.434	0.819	0.521
$f_{PSSM} + f_{OBV} + f_{SS}$	0.337	0.975	0.431	0.830	0.517
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.419	0.958	0.450	0.832	0.535
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.439	0.952	0.453	0.836	0.539
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.442	0.949	0.445	0.832	0.535

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

To further investigate the performance of $\mathbb{JSD}$-based features proposed in this study, we analyzed two additional datasets, namely RBscore [2] and PreDNA datasets [37]. Although the RBscore and PreDNA datasets initially contain 381 and 224 DNA-binding proteins, respectively, we have eliminated a few proteins since they are either included in our training dataset or ineligible due to their MSAs. Consequently, we constructed RF classifier using 263 proteins (which were also used for cross-validation) and randomly selecting 60 proteins from each dataset for testing, respectively. The results of these analyses consistently suggest that our new features show great complementary effect to the previous features which often leads to clear improvement of the classification performance (see Tables 3 and 4). The detailed performance of classifier on different features using different cut-offs for each dataset can be found in Appendix A (see Tables A3–A6).

Considering the AUC-ROC and AUC-PR as the only evaluation factor, results indicate that the RF classifier often achieved its best performance based on both cut-off distances if we combine our new $f_{\mathbb{JSD}\text{-t}}$ feature together with the existing three features (see Tables 1–3). Interestingly, by analyzing the PreDNA dataset we observed that RF classifier with $f_{\mathbb{JSD}}$ or $f_{\mathbb{JSD}\text{-t}}$ features for the cut-off of 3.5 Å showed similar performance. However, regarding to the distance cut-off of 5 Å, the classifier with $f_{\mathbb{JSD}}$ feature reached slightly better performance than those with $f_{\mathbb{JSD}\text{-t}}$ feature (see Table 4). After looking at the overall performances, it is inferred that adding our new features can boost the performance of the RF classifier in terms of AUC-ROC and AUC-PR.

Table 3. Prediction performance of Random Forest (RF) classifier on RBscore dataset using different distance cut-offs.

Cut-Off	Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
3.5 Å	$f_{PSSM} + f_{OBV} + f_{SS}$	0.517	0.976	0.534	0.896	0.528
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.58	0.967	0.54	0.907	0.543
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.612	0.963	0.546	0.910	0.551
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.601	0.962	0.531	0.909	0.546
5.0 Å	$f_{PSSM} + f_{OBV} + f_{SS}$	0.499	0.98	0.584	0.895	0.641
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.57	0.968	0.595	0.908	0.661
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.592	0.965	0.60	0.908	0.665
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.594	0.964	0.597	0.907	0.663

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

Table 4. Prediction performance of RF classifier on PreDNA dataset using different distance cut-offs.

Cut-Off	Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
3.5 Å	$f_{PSSM} + f_{OBV} + f_{SS}$	0.428	0.977	0.458	0.867	0.451
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.511	0.97	0.488	0.885	0.488
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.539	0.962	0.475	0.888	0.488
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.539	0.961	0.47	0.886	0.488
5.0 Å	$f_{PSSM} + f_{OBV} + f_{SS}$	0.395	0.98	0.488	0.858	0.530
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.48	0.968	0.511	0.874	0.563
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.506	0.962	0.51	0.873	0.560
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.499	0.96	0.498	0.871	0.555

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

2.2. Position Analysis of the MYC-MAX Protein

The proto-oncobenic transcription factor MYC-MAX (PDB-Entry 1NKP) is a heterodimer protein complex that is active in cell proliferation and is over-expressed in many different cancer types [38]. MYC-MAX transcription factors bind to Enhancer boxes (a core element of the promoter that consists of six nucleotides) and activate transcription of the underlying genes [39].

The amino acid chain of MYC protein consists of 88 residues, ten of which are known DNA-binding sites indicating that their distances to DNA are less than 3.5 Å. Applying RF classifier, which takes a majority vote among the random tree classifiers, with our first feature ($f_{\mathbb{JSD}}$) combined with existing features, we predicted in total 17 residue positions to be DNA-binding in MYC protein. Seven out of these positions (H906, N907, E910, R913, R914, P938, K939) correspond to the true DNA-binding sites of this protein. While the sites R913, R914, P938, and K939 could also be identified by RF classifier without using our new $\mathbb{JSD}$-based features, the remaining three binding sites could only be detected using our features (for details see Table 5 and Figure 1). Interestingly, using $f_{\mathbb{JSD}\text{-t}}$ together with f_{PSSM}, f_{OBV}, and f_{SS}, the RF classifier correctly predicted these seven positions again as binding sites.

The second protein in the proto-oncobenic transcription factor complex is the MAX protein which consists of 83 residues including nine DNA-binding sites. Using $f_{\mathbb{JSD}}$ or $f_{\mathbb{JSD}\text{-t}}$ together with existing features individually, we observed 14 and 13 residue positions to be DNA-binding in MAX protein, respectively. Eight of the predicted positions (H207, N208, E211, R212, R214, R215, S238, R239) found by using either of our both features are true DNA-binding sites in MAX protein. However, without using our new features the RF classifier could only identify two (S238, R239) out of nine true DNA-binding sites in MAX protein (for details see Table 5 and Figure 1). Further, we observed that, the usage of $f_{\mathbb{JSD}\text{-t}}$ leads to the reduction of false positive predictions in identifying DNA-binding sites in MAX protein.

Table 5. Prediction performance of RF classifier on different features using a cut-off of 3.5 Å for MYC-MAX protein complex (Protein Data Bank (PDB)-Entry 1NKP).

Protein	Feature	Sensitivity	Specificity	MCC
MYC	$f_{PSSM} + f_{OBV} + f_{SS}$	0.30	0.941	0.282
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.70	0.853	0.448
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.70	0.853	0.448
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.70	0.868	0.470
MAX	$f_{PSSM} + f_{OBV} + f_{SS}$	0.222	1.0	0.447
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.888	0.906	0.664
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.888	0.922	0.697
	$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.889	0.922	0.697

MCC: Matthews correlation coefficient.

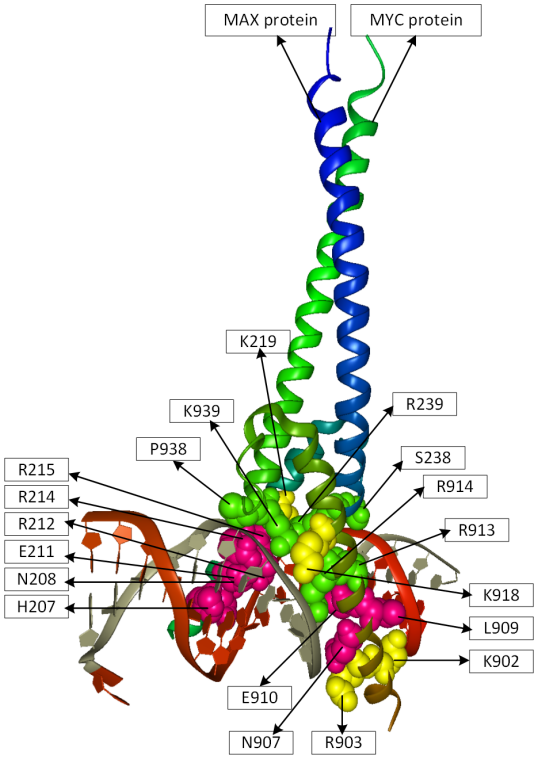

Figure 1. DNA-binding sites in proto-oncobenic transcription factor MYC-MAX protein complex (PDB-Entry 1NKP). Green spheres denote positions of the DNA-binding sites in both proteins which are detected by RF classifier either using the existing features (f_{PSSM}, f_{OBV}, and f_{SS}) alone or combining our new features with these existing features together. Purple spheres show the localization of additional binding sites which were only found by RF classifier using our new features with existing features. Moreover, there are further three binding sites in MYC protein and one binding site in MAX protein, shown with yellow spheres, that could not be identified by the classifier.

Moreover, when statistically evaluating both of our features, we observed that using our sequence-based features RF classifier reaches a significantly improved performance in identifying DNA-binding sites of both proteins with significantly higher sensitivity and MCC values whereas the specificity is moderately decreased. The simultaneous usage of both of our features together with f_{PSSM}, f_{OBV}, and f_{SS} could result in the decrement of specificity or MCC values. The details are presented in Table 5.

3. Materials and Methods

In this section, we describe in particular the data we have used and our new residue-wise features designed to predict DNA-binding sites in proteins.

3.1. Materials

To compile our data needed for training and test, we started with the DBP-374 data set of representative protein-DNA complexes from the Protein Data Bank (PDB) [40] published by Wu et al. [5]. Having performed a comparison with the new PDB version, we calculate for every remaining protein a multiple sequence alignment (MSA) using HHblits and the UniProt20 database (version from June 2015) [41]. We eliminated all proteins, the MSA of which has less than 125 rows, so that we finally ended up with a dataset of 263 protein-DNA complexes and associated MSAs. To obtain our results we perform a five-fold cross validation.

As in [5], an amino acid residue is regarded as a binding site, if it contains at least one atom at distance of less than or equal to 3.5 Å or 5 Å from any atom of DNA molecule in the DNA-protein complex. Otherwise it is treated as non-binding site. For the distance cut-off of 3.5 Å, our set contains 4298 binding sites and 44,805 non-binding sites. For the distance cut-off of 5 Å, however, our data set contains 7211 binding sites and 41,892 non-binding sites.

3.2. Methods

Let M be a multiple sequence alignment, where its first row represents the protein under study. Every residue of that protein is then uniquely determined by its column. In what follows, we identify the residues of the protein with their columns of the MSA.

Grosse et al. [35] pointed out that the Jensen–Shannon divergence ($\mathbb{JSD}$) is extremely useful when it comes to discriminate between two (or more) sources. Capra and Singh [34] carefully discussed several information theoretic measures like Shannon entropy, von Neumann entropy, relative entropy, and sum-of-pair measures to assess sequence conservation. They were the first using $\mathbb{JSD}$ in this context and stated its superiority. Gültas et al. [32] showed that the Jensen–Shannon divergence in the context of quantum information theory is of remarkable power. These three articles encouraged us to use $\mathbb{JSD}$ in this study. Our first idea is to design a new feature for the prediction of DNA-binding sites in proteins which leverages the *Jensen–Shannon divergence*

$$\mathbb{JSD}\left(\mathbf{p}_k \,\|\mathbf{p}_{nd}\right) := \mathbb{H}\left(\left(\mathbf{p}_k + \mathbf{p}_{nd}\right)/2\right) - \left(\mathbb{H}\left(\mathbf{p}_k\right) + \mathbb{H}\left(\mathbf{p}_{nd}\right)\right)/2. \tag{1}$$

Therein, $\mathbf{p}_k$ is the empirical amino acid distribution of the k-th column of the query MSA M, and $\mathbf{p}_{nd}$ is the *null distribution* taken over all non-binding sites of our training data.

More precisely, we represent every column k of every MSA M considered by a 20×20 counting matrix $C\left(M_{\bullet k}\right)$. The matrix C is symmetric and its rows as well as columns are indexed by the 20 amino acids. For every ordered pair of amino acids $\left(a, a'\right)$, the matrix coefficient $C\left(M_{\bullet k}\right)_{aa'}$ is equal to the number of ordered pairs (i, j) $(i \neq j)$ of row indices of M such that $M_{ik} = a$ and $M_{jk} = a'$.

To compute the null distribution $\mathbf{p}_{nd}$, we first set up the 20×20 counting matrix $\mathcal{C}_{nd}$ using our training data. $\mathcal{C}_{nd}$ is the sum over all matrices $C\left(M_{\bullet k}\right)$, where M ranges over all training MSAs and k ranges over all non-binding site columns of M. Next, the rows of $\mathcal{C}_{nd}$ are added up. Finally, the resulting row vector is normalized to obtain $\mathbf{p}_{nd}$.

There is nothing wrong with the idea that a large value $\mathbb{JSD}\left(\mathbf{p}_k \,\|\mathbf{p}_{nd}\right)$ indicates that k is a DNA-binding residue. However, no information on binding sites is integrated. Only the non-binding sites of our training data are used to compute $\mathbf{p}_{nd}$. As we have seen in [32] and [36], transforming empirical amino acid distributions of MSA columns by a carefully designed doubly stochastic matrix is an effective way to integrate the binding site signals. To this end, we first set up a counting matrix $\mathcal{C}_{bind}$ in a way similar to that of calculating the matrix $\mathcal{C}_{nd}$. The difference is that the variable column index k now ranges over all binding site columns of the training MSAs. Taking the counting matrix $\mathcal{C}_{bind}$ as input, the doubly stochastic matrix $\mathcal{D}$ is computed by means of the canonical row-column normalization procedure [42].

Let M be the query MSA having ℓ columns. Compared with [32] and [36], we enhance the effect of transforming M's empirical column distributions by means of the doubly stochastic matrix $\mathcal{D}$ just defined. Let k be a column index of M. First, we compute the matrix product $C^{(t)}\left(M_{\bullet k}\right) := C\left(M_{\bullet k}\right) \cdot \mathcal{D}$. Second, we add up all of $C^{(t)}\left(M_{\bullet k}\right)$'s rows. Finally, we normalize the resulting row to obtain the transformed empirical row distribution $\mathbf{p}_k^{(t)}$.

We define two *window scores* $\text{score}_{\mathbb{JSD},M}\left(k\right)$ and $\text{score}_{\mathbb{JSD}\text{-t},M}\left(k\right)$ of residue k w.r.t. query MSA M, where the window $\mathfrak{w}(k)$ surrounding k formally equals $\{k-3, k-2, k-1, k, k+1, k+2, k+3\} \cap \{1, 2, \ldots, \ell\}$. Clearly, if $k \in \{4, 5, \ldots, \ell-3\}$, $|\mathfrak{w}(k)| = 7$. Otherwise $|\mathfrak{w}(k)| \in \{4, 5, 6\}$. Recapitulate that for any real x the binomial coefficient $\binom{x}{2}$ equals $x(x-1)/2$. We define the scores as follows.

$$\text{score}_{\mathbb{JSD},M}(k) := \frac{\sum_{\iota \in \mathfrak{w}(k)} (4 - |k - \iota|)\, \mathbb{JSD}(\mathbf{p}_{k+\iota} \,\|\, \mathbf{p}_{nd})}{16 - \binom{8 - |\mathfrak{w}(k)|}{2}} \tag{2}$$

$$\text{score}_{\mathbb{JSD}\text{-t},M}(k) := \frac{\sum_{\iota \in \mathfrak{w}(k)} (4 - |k - \iota|)\, \mathbb{JSD}\left(\mathbf{p}^{(t)}_{k+\iota} \,\|\, \mathbf{p}_{nd}\right)}{16 - \binom{8 - |\mathfrak{w}(k)|}{2}} \tag{3}$$

The preceding two score definitions are motivated as follows. Bartlett et al. [43] and Panchenko et al. [44] pointed out that exploiting conservation properties of spatial neighbors is useful to predict a residue as functionally important. Since the 3D structures are often unavailable, Capra and Singh [34] developed a window score for such predictions. The concrete shape of our scores takes pattern form Janda et al. [45], who in turn refer to Fischer et al. [33]. Our scores are convex combinations of the Jensen–Shannon terms associated with the residues belonging to the surrounding window $\mathfrak{w}(k)$. The weights fall linearly in the distance from k.

In a last step, we transform two window scores according to Equations (2) and (3) with respect to the query MSA M into final scores using the Equations (4) and (5), respectively. To this end, for every column index $k \in \{1, 2, \ldots, \ell\}$ of M we define:

$$\mathrm{f}_{\mathbb{JSD},M}(k) := \frac{\left|\{k' \mid 1 \le k' \le \ell,\ \text{score}_{\mathbb{JSD},M}(k) \ge \text{score}_{\mathbb{JSD},M}(k')\}\right|}{\ell} \tag{4}$$

$$\mathrm{f}_{\mathbb{JSD}\text{-t},M}(k) := \frac{\left|\{k' \mid 1 \le k' \le \ell,\ \text{score}_{\mathbb{JSD}\text{-t},M}(k) \ge \text{score}_{\mathbb{JSD}\text{-t},M}(k')\}\right|}{\ell}. \tag{5}$$

The Equations (4) and (5) are basically the determination of the percentage of scores below the current one at index k. This transformation procedure is essential because it converts MSA-dependent window scores to MSA-independent scores.

To demonstrate the benefit of our new features, we adopt the features f_{PSSM}, f_{OBV} and f_{SS} devised in [5]. Together with our two new features $\mathrm{f}_{\mathbb{JSD}}$ and $\mathrm{f}_{\mathbb{JSD}\text{-t}}$, we plugged them into the Random Forest (RF) classifier [46] (see Tables 1 and 2 for the combinations we used). For the RF implementation we used the WEKA data mining software [47].

To deal with the imbalanced data problem, we applied bagging techniques suggested in [48]. Since we make use of five-fold cross validation, we randomly split the dataset into 5 roughly equal-sized parts. Every training phase performed on 4 parts consists of 11 sub-phases. In each such sub-phase we randomly draw twice as many non-binding sites as there are binding sites. We then construct a Random Forest (RF) taking those non-binding sites and all binding sites of the 4 parts as input. Finally, for each instance of the validation part the majority vote of above 11 RF classifiers was taken.

4. Discussion

Our results show that combining either feature $\mathrm{f}_{\mathbb{JSD}\text{-t}}$ or feature $\mathrm{f}_{\mathbb{JSD}}$ with the three features f_{PSSM}, f_{OBV} and f_{SS} we have adopted from [5] clearly boosts the performance of the RF-based classifier in identifying the DNA-binding sites in proteins, where feature $\mathrm{f}_{\mathbb{JSD}\text{-t}}$ generally reaches a slightly better performance than feature $\mathrm{f}_{\mathbb{JSD}}$.

Although our two new features and PSSMs are derived from MSAs, Tables 1 and 2 clearly demonstrate that these approaches carry distinct information. Thus they capture different kinds of evolutionary information. The reason for this essential difference can be explained based on the underlying algorithms. While the PSSM approach consists of statistic which indicates how likely a certain amino acids occurs at a certain position, our $\mathbb{JSD}$-based approach measures the divergence of a certain distribution to a known non-binding site distribution.

The superiority of feature $f_{\mathbb{JSD}\text{-t}}$ to feature $f_{\mathbb{JSD}}$ deserves an explanation attempt. Feature $f_{\mathbb{JSD}}$ does not integrate any information on DNA-binding sites. Only training non-binding sites are used. In contrast, feature $f_{\mathbb{JSD}\text{-t}}$ additionally uses a doubly stochastic matrix gained from the training binding sites. The effect on empirical amino acid column distributions of the transformation we have devised using that matrix is the following. The empirical column probabilities of amino acids are merged, if it is very likely to co-observe them in a binding site column. Since the amino acid content of binding site columns and non-binding site columns differ, the distance between $f_{\mathbb{JSD}\text{-t},M}(k)$ and $f_{\mathbb{JSD}\text{-t},M}(k')$ is larger and more significant than the distance between $f_{\mathbb{JSD},M}(k)$ and $f_{\mathbb{JSD},M}(k')$, where k is a binding site column of MSA M, and k' is a non-binding site column.

At first glance it is surprising that adding both feature $f_{\mathbb{JSD}\text{-t}}$ and feature $f_{\mathbb{JSD}}$ to the feature triplet $(f_{PSSM}, f_{OBV}, f_{SS})$ is worse than adding feature $f_{\mathbb{JSD}\text{-t}}$ alone. Taking into account what we have mentioned in the preceding paragraph, it turns out that if feature $f_{\mathbb{JSD}\text{-t}}$ is already there, feature $f_{\mathbb{JSD}}$ may increase the noise.

5. Conclusions

In this work, we report a new sequence-based feature extraction method for the identification of DNA binding sites in proteins. For this purpose, we adopt the ideas from Capra et al. [34] and our previous studies CMF [36] and QCMF [32]. Our approach is an information theoretic method that applies the Jensen–Shannon divergence ($\mathbb{JSD}$) for amino acid distributions of each site in a protein in two different ways. First, the $\mathbb{JSD}$ is applied to quantify the differences between observed amino acid distributions of sites and the background distribution of non-binding sites. Second, we transform the observed distributions of sites through a doubly stochastic matrix to incorporate biochemical signals of binding residues in the calculation of $\mathbb{JSD}$ that results in the intensification of the DNA-binding residue signals from the non-binding signals. The results of our study show that the additional usage of our new features ($f_{\mathbb{JSD}\text{-t}}$ or feature $f_{\mathbb{JSD}}$) in combination with existing features is significantly boosts the performance of RF classifier in identifying DNA binding sites in proteins. Our results further indicate the importance of our second feature ($f_{\mathbb{JSD}\text{-t}}$) since taking into account the binding site signals in the calculation of $\mathbb{JSD}$ metric, the characteristics of DNA binding residues are enhanced. As a consequence, an intensification of the signal caused by DNA binding sites from non-binding sites occurs and thus the classifier achieves its improved performance.

Acknowledgments: We thank our colleagues Edgar Wingender, Martin Haubrock and Sebastian Zeidler for their helpful advice and insights at early stages of this project. We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

Author Contributions: Mehmet Gültas developed the model. Stephan Waack adjusted the model together with Mehmet Gültas. Truong Khanh Linh Dang developed the model together with Mehmet Gültas, designed and implemented the tool and interpreted the results together with Cornelia Meckbach, Rebecca Tacke and Mehmet Gültas. Cornelia Meckbach and Rebecca Tacke studied the DNA binding sites in MYC-MAX protein complex. Mehmet Gültas conceived of and managed the project and wrote the final version of the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The detailed performance of the RF classifier on different features using different cut-offs for RBscore and PreDNA datasets.

Appendix A.1. Performance Measures with Standard Error

Table A1. Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å. The prediction system was evaluated by five-fold cross validation.

Feature	Sensitivity ± SE(%)	Specificity ± SE(%)	MCC ± SE(%)
f_{PSSM}	29.2 ± 2.20	96.3 ± 0.46	30.7 ± 0.95
f_{PSSM} + $f_{\mathbb{JSD}}$	38.5 ± 3.04	94.9 ± 0.57	34.9 ± 1.7
f_{PSSM} + $f_{\mathbb{JSD}\text{-t}}$	41.0 ± 3.23	93.9 ± 0.57	35.0 ± 1.85
f_{PSSM} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	41.4 ± 3.42	94.0 ± 0.51	34.8 ± 2.07
f_{PSSM} + f_{SS}	33.9 ± 2.32	95.8 ± 0.37	33.4 ± 1.36
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$	41.6 ± 3.05	95.0 ± 0.46	37.8 ± 2.19
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	44.1 ± 3.12	94.0 ± 0.43	37.2 ± 2.37
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	43.9 ± 3.14	94.0 ± 0.40	37.0 ± 2.25
f_{PSSM} + f_{OBV} + f_{SS}	36.7 ± 2.07	96.8 ± 0.27	39.8 ± 1.58
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$	42.2 ± 2.70	95.8 ± 0.42	40.9 ± 1.95
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	44.7 ± 3.05	95.0 ± 0.38	40.3 ± 1.98
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	44.4 ± 3.12	94.7 ± 0.39	39.3 ± 2.02

Table A2. Prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å. The prediction system was evaluated by five-folds cross validation.

Feature	Sensitivity ± SE(%)	Specificity ± SE(%)	MCC ± SE(%)
f_{PSSM}	28.6 ± 2.56	96.6 ± 0.47	35.0 ± 1.43 5
f_{PSSM} + $f_{\mathbb{JSD}}$	39.5 ± 2.89	95.0 ± 0.55	40.7 ± 1.99
f_{PSSM} + $f_{\mathbb{JSD}\text{-t}}$	41.8 ± 3.02	94.3 ± 0.62	41.1 ± 2.05
f_{PSSM} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	42.6 ± 3.25	94.2 ± 0.54	41.4 ± 2.37
f_{PSSM} + f_{SS}	33.4 ± 2.34	96.3 ± 0.38	38.6 ± 1.90
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$	42.4 ± 2.97	95.1 ± 0.61	43.6 ± 2.43
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	44.8 ± 2.99	94.4 ± 0.56	43.8 ± 2.45
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	44.5 ± 3.04	94.4 ± 0.50	43.4 ± 2.35
f_{PSSM} + f_{OBV} + f_{SS}	33.7 ± 2.48	97.5 ± 0.35	43.1 ± 2.05
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$	41.9 ± 2.89	95.8 ± 0.55	45.0 ± 2.39
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	43.9 ± 2.89	95.2 ± 0.48	45.3 ± 2.32
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	44.2 ± 2.91	94.9 ± 0.54	44.5 ± 2.24

Appendix A.2. RBscore Dataset Analysis

Table A3. The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f_{PSSM}	0.458	0.974	0.476	0.866	0.460
f_{PSSM} + $f_{\mathbb{JSD}}$	0.56	0.965	0.514	0.894	0.518
f_{PSSM} + $f_{\mathbb{JSD}\text{-t}}$	0.597	0.957	0.511	0.899	0.523
f_{PSSM} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	0.591	0.958	0.511	0.90	0.526
f_{PSSM} + f_{SS}	0.512	0.97	0.501	0.878	0.476
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$	0.581	0.96	0.511	0.899	0.520
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	0.611	0.953	0.508	0.903	0.526
f_{PSSM} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	0.613	0.953	0.509	0.902	0.528
f_{PSSM} + f_{OBV} + f_{SS}	0.517	0.976	0.534	0.896	0.528
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$	0.58	0.967	0.54	0.907	0.543
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}\text{-t}}$	0.612	0.963	0.546	0.910	0.551
f_{PSSM} + f_{OBV} + f_{SS} + $f_{\mathbb{JSD}}$ + $f_{\mathbb{JSD}\text{-t}}$	0.601	0.962	0.531	0.909	0.546

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

Table A4. The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f_{PSSM}	0.445	0.977	0.528	0.873	0.589
$f_{PSSM} + f_{\mathbb{JSD}}$	0.553	0.968	0.579	0.899	0.643
$f_{PSSM} + f_{\mathbb{JSD}\text{-t}}$	0.57	0.962	0.572	0.900	0.642
$f_{PSSM} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.569	0.963	0.574	0.895	0.642
$f_{PSSM} + f_{SS}$	0.49	0.973	0.547	0.880	0.602
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}}$	0.578	0.963	0.583	0.902	0.648
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.605	0.958	0.587	0.904	0.652
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.603	0.959	0.587	0.902	0.653
$f_{PSSM} + f_{OBV} + f_{SS}$	0.499	0.98	0.584	0.895	0.641
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.57	0.968	0.595	0.908	0.661
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.592	0.965	0.60	0.908	0.665
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.594	0.964	0.597	0.907	0.663

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

Appendix A.3. PreDNA Dataset Analysis

Table A5. The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 3.5 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f_{PSSM}	0.378	0.977	0.41	0.840	0.391
$f_{PSSM} + f_{\mathbb{JSD}}$	0.498	0.963	0.448	0.865	0.453
$f_{PSSM} + f_{\mathbb{JSD}\text{-t}}$	0.543	0.953	0.445	0.869	0.451
$f_{PSSM} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.538	0.956	0.453	0.869	0.455
$f_{PSSM} + f_{SS}$	0.393	0.975	0.417	0.847	0.402
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}}$	0.501	0.966	0.461	0.872	0.463
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.545	0.959	0.465	0.876	0.468
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.523	0.958	0.449	0.875	0.465
$f_{PSSM} + f_{OBV} + f_{SS}$	0.428	0.977	0.458	0.867	0.451
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.511	0.97	0.488	0.885	0.488
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.539	0.962	0.475	0.888	0.488
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.539	0.961	0.47	0.886	0.488

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

Table A6. The detailed prediction performance of Random Forest (RF) classifier on different features using a cut-off of 5.0 Å.

Feature	Sensitivity	Specificity	MCC	AUC-ROC	AUC-PR
f_{PSSM}	0.373	0.979	0.463	0.833	0.496
$f_{PSSM} + f_{\mathbb{JSD}}$	0.485	0.962	0.495	0.858	0.540
$f_{PSSM} + f_{\mathbb{JSD}\text{-t}}$	0.496	0.953	0.475	0.858	0.534
$f_{PSSM} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.495	0.955	0.479	0.857	0.535
$f_{PSSM} + f_{SS}$	0.389	0.977	0.47	0.839	0.501
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}}$	0.49	0.963	0.501	0.863	0.550
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.503	0.957	0.492	0.865	0.547
$f_{PSSM} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.504	0.958	0.497	0.865	0.550
$f_{PSSM} + f_{OBV} + f_{SS}$	0.395	0.98	0.488	0.858	0.530
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}}$	0.48	0.968	0.511	0.874	0.563
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}\text{-t}}$	0.506	0.962	0.51	0.873	0.560
$f_{PSSM} + f_{OBV} + f_{SS} + f_{\mathbb{JSD}} + f_{\mathbb{JSD}\text{-t}}$	0.499	0.96	0.498	0.871	0.555

MCC: Matthews correlation coefficient; AUC-ROC: area under the receiver operating characteristics (ROC) curve; AUC-PR: area under the precision-recall curve.

References

1. Liu, B.; Wang, S.; Wang, X. DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Sci. Rep.* **2015**, *5*, 15479.
2. Miao, Z.; Westhof, E. Prediction of nucleic acid binding probability in proteins: A neighboring residue network based score. *Nucleic Acids Res.* **2015**, *43*, 5340–5351.
3. Si, J.; Zhang, Z.; Lin, B.; Schroeder, M.; Huang, B. MetaDBSite: A meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.* **2011**, *5* (Suppl. S1), S7.
4. Ma, X.; Guo, J.; Liu, H.D.; Xie, J.M.; Sun, X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1766–1775.
5. Wu, J.; Liu, H.; Duan, X.; Ding, Y.; Wu, H.; Bai, Y.; Sun, X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* **2009**, *25*, 30–35.
6. Liu, B.; Xu, J.; Fan, S.; Xu, R.; Zhou, J.; Wang, X. PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. *Mol. Inform.* **2015**, *34*, 8–17.
7. Xu, R.; Zhou, J.; Wang, H.; He, Y.; Wang, X.; Liu, B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* **2015**, *9* (Suppl. S1), S10.
8. Dong, Q.; Wang, S.; Wang, K.; Liu, X.; Liu, B. Identification of DNA-binding proteins by auto-cross covariance transformation. In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015; pp. 470–475.
9. Wei, L.; Tang, J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2016**, in press.
10. Waris, M.; Ahmad, K.; Kabir, M.; Hayat, M. Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. *Neurocomputing* **2016**, *199*, 154–162.
11. Zhou, J.; Xu, R.; He, Y.; Lu, Q.; Wang, H.; Kong, B. PDNAsite: Identification of DNA-binding Site from Protein Sequence by Incorporating Spatial and Sequence Context. *Sci. Rep.* **2016**, *6*, 27653.
12. Jones, S.; Shanahan, H.P.; Berman, H.M.; Thornton, J.M. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* **2003**, *31*, 7189–7198.
13. Ahmad, S.; Gromiha, M.M.; Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* **2004**, *20*, 477–486.
14. Bhardwaj, N.; Langlois, R.E.; Zhao, G.; Lu, H. Structure based prediction of binding residues on DNA-binding proteins. In Proceedings of the IEEE 27th Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS 2005), Shanghai, China, 1–4 September 2005; pp. 2611–2614.
15. Ahmad, S.; Sarai, A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinform.* **2005**, *6*, 33.
16. Kuznetsov, I.B.; Gou, Z.; Li, R.; Hwang, S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins* **2006**, *64*, 19–27.
17. Wang, L.; Brown, S.J. Prediction of DNA-binding residues from sequence features. *J. Bioinform. Comput. Biol.* **2006**, *4*, 1141–1158.
18. Wang, L.; Brown, S.J. BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* **2006**, *34*, W243–W248.
19. Ofran, Y.; Mysore, V.; Rost, B. Prediction of DNA-binding residues from sequence. *Bioinformatics* **2007**, *23*, i347–i353.
20. Siggers, T.W.; Honig, B. Structure-based prediction of C2H2 zinc-finger binding specificity: Sensitivity to docking geometry. *Nucleic Acids Res.* **2007**, *35*, 1085–1097.
21. Tjong, H.; Zhou, H.X. DISPLAR: An accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.* **2007**, *35*, 1465–1477.
22. Nimrod, G.; Schushan, M.; Szilágyi, A.; Leslie, C.; Ben-Tal, N. iDBPs: A web server for the identification of DNA binding proteins. *Bioinformatics* **2010**, *26*, 692–693.
23. Wang, L.; Huang, C.; Yang, M.Q.; Yang, J.Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.* **2010**, *4* (Suppl. S1), S3.
24. Miao, Z.; Westhof, E. A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs. *PLoS Comput. Biol.* **2015**, *11*, e1004639.
25. Yan, J.; Friedrich, S.; Kurgan, L. A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. *Brief. Bioinform.* **2015**, *17*, 88–105.

26. Yan, C.; Terribilini, M.; Wu, F.; Jernigan, R.L.; Dobbs, D.; Honavar, V. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinform.* **2006**, *7*, 262.
27. Hwang, S.; Gou, Z.; Kuznetsov, I.B. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* **2007**, *23*, 634–636.
28. Huang, Y.F.; Huang, C.C.; Liu, Y.C.; Oyang, Y.J.; Huang, C.K. DNA-binding residues and binding mode prediction with binding-mechanism concerned models. *BMC Genom.* **2009**, *10* (Suppl. S3), S23.
29. Wong, K.C.; Li, Y.; Peng, C.; Moses, A.M.; Zhang, Z. Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res.* **2015**, *43*, 10180–10189.
30. Wang, L.; Yang, M.Q.; Yang, J.Y. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genom.* **2009**, *10* (Suppl. S1), S1.
31. Eggeling, R.; Roos, T.; Myllymäki, P.; Grosse, I. Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinform.* **2015**, *16*, doi:10.1186/s12859-015-0797-4.
32. Gültas, M.; Düzgün, G.; Herzog, S.; Jäger, S.J.; Meckbach, C.; Wingender, E.; Waack, S. Quantum coupled mutation finder: Predicting functionally or structurally important sites in proteins using quantum Jensen–Shannon divergence and CUDA programming. *BMC Bioinform.* **2014**, *15*, 96.
33. Fischer, J.; Mayer, C.E.; Söding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* **2008**, *24*, 613–620.
34. Capra, J.A.; Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **2007**, *23*, 1875–1882.
35. Grosse, I.; Bernaola-Galván, P.; Carpena, P.; Román-Roldán, R.; Oliver, J.; Stanley, H.E. Analysis of symbolic sequences using the Jensen–Shannon divergence. *Phys. Rev. E* **2002**, *65*, 041905.
36. Gültas, M.; Haubrock, M.; Tüysüz, N.; Waack, S. Coupled mutation finder: A new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. *BMC Bioinform.* **2012**, *13*, 225.
37. Li, T.; Li, Q.Z.; Liu, S.; Fan, G.L.; Zuo, Y.C.; Peng, Y. PreDNA: Accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. *Bioinformatics* **2013**, *29*, 678–685,
38. Krall, A.; Brunn, J.; Kankanala, S.; Peters, M.H. A simple contact mapping algorithm for identifying potential peptide mimetics in protein–protein interaction partners. *Proteins* **2014**, *82*, 2253–2262.
39. Nair, S.K.; Burley, S.K. X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors. *Cell* **2003**, *112*, 193–205.
40. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
41. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173–175.
42. Cappellini, V.; Sommer, H.J.; Bruzda, W.; Zyczkowski, K. Random bistochastic matrices. *J. Phys. A Math. Theor.* **2009**, *42*, 36.
43. Bartlett, G.J.; Porter, C.T.; Borkakoti, N.; Thornton, J.M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **2002**, *324*, 105–121.
44. Panchenko, A.R.; Kondrashov, F.; Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **2004**, *13*, 884–892.
45. Janda, J.O.; Busch, M.; Kück, F.; Porfenenko, M.; Merkl, R. CLIPS-1D: Analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinform.* **2012**, *13*, 55.
46. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
47. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18.
48. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.