Analytical Biochemistry 612 (2021) 113954

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Single-stranded and double-stranded DNA-binding protein prediction using HMM profiles

Ronesh Sharma [a,*,2], Shiu Kumar [a,2], Tatsuhiko Tsunoda [b,c,d], Thirumananseri Kumarevel [e,1], Alok Sharma [b,c,f,g,1]

[a] School of Electrical and Electronics Engineering, Fiji National University, Suva, Fiji
[b] Laboratory of Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Japan
[c] Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo, 113-8510, Japan
[d] Laboratory of Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo, 113-0033, Japan
[e] Laboratory for Transcription Structural Biology, RIKEN Center for Biosystems Dynamics Research, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan
[f] School of Engineering and Physics, The University of the South Pacific, Suva, Fiji
[g] Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD, Australia

ARTICLE INFO

Keywords:
DNA-Binding proteins
SSBs
DSBs
Hidden markov model
Support vector machine
Random forest
K-nearest neighbors

ABSTRACT

Background: DNA-binding proteins perform important roles in cellular processes and are involved in many biological activities. These proteins include crucial protein-DNA binding domains and can interact with single-stranded or double-stranded DNA, and accordingly classified as single-stranded DNA-binding proteins (SSBs) or double-stranded DNA-binding proteins (DSBs). Computational prediction of SSBs and DSBs helps in annotating protein functions and understanding of protein-binding domains.
Results: Performance is reported using the DNA-binding protein dataset that was recently introduced by Wang et al., [1]. The proposed method achieved a sensitivity of 0.600, specificity of 0.792, AUC of 0.758, MCC of 0.369, accuracy of 0.744, and F-measure of 0.536, on the independent test set.
Conclusion: The proposed method with the hidden Markov model (HMM) profiles for feature extraction, outperformed the benchmark method in the literature and achieved an overall improvement of approximately 3%. The source code and supplementary information of the proposed method is available at https://github.com/roneshsharma/Predict-DNA-binding-proteins/wiki.

1. Background

DNA-binding proteins are involved in a variety of biological processes, such as DNA repair, DNA packing, viral infection and DNA replication [1–5]. Identification of these proteins is a step towards annotating the protein functions and understanding the binding specificity [6]. Although many experimentally determined protein-DNA structures are deposited into the protein databank (PDB), only a small portion is listed in comparison with the protein-DNA complexes present in nature [6,7]. A large number of DNAs and protein sequences has been generated, of which many are DNA-binding proteins. DNA-binding proteins can be identified using various biological experiments, but it is expensive and time-consuming. In this respect, it is highly desirable to design computational methods to determine the DNA-binding proteins. Computational methods are applied in two categories; first is the use of protein structure information and second, is the use of sequence information only. Incorporating protein structures to predict the DNA-binding proteins exhibits improved performance. However, these structures are not always available [1]. Thus, the sequence information is only used for prediction.

Compared with the biological experiments; computational methods are gaining momentum as it is inexpensive and reliable to identify protein functions. In recent years, many efforts have been made whereby machine learning approaches are used for DNA-binding protein

[*] Corresponding author.
E-mail addresses: sharmaronesh@yahoo.com (R. Sharma), shiu748@gmail.com (S. Kumar), tsunoda@bs.s.u-tokyo.ac.jp (T. Tsunoda), kumarevel.thirumananseri@riken.jp (T. Kumarevel), alok.fj@gmail.com (A. Sharma).
[1] Last Authors.
[2] Equal contribution.

https://doi.org/10.1016/j.ab.2020.113954
Received 16 April 2020; Received in revised form 26 August 2020; Accepted 10 September 2020
Available online 15 September 2020

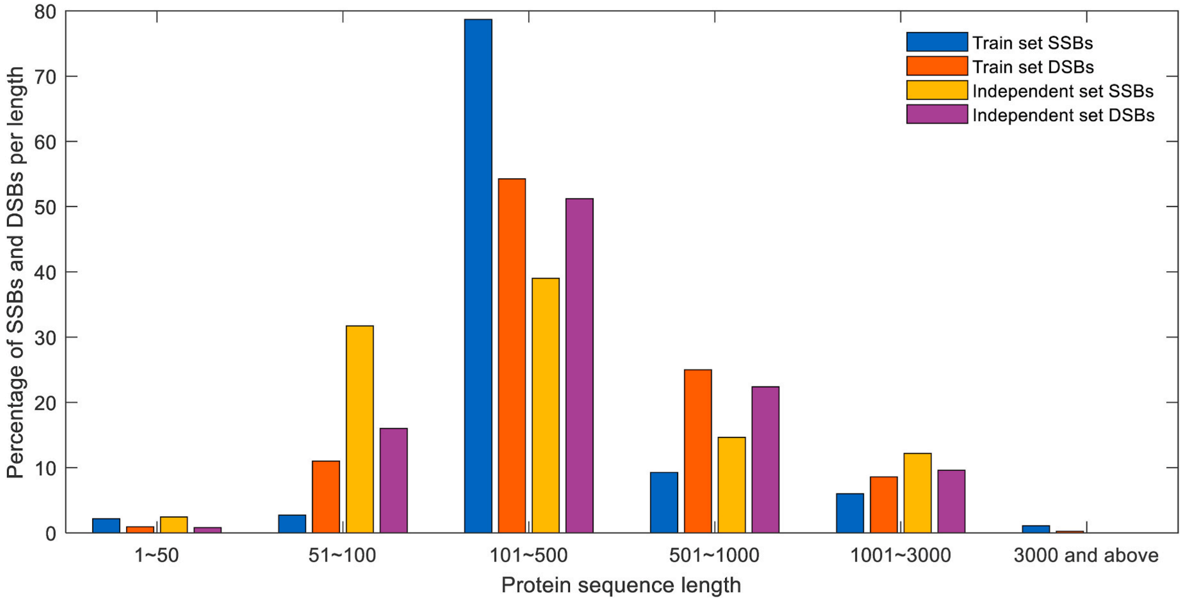

Fig. 1. Length distribution of the SSBs and DSBs in train and independent sets.

prediction [8–11]. These approaches involve feature extraction and classifier development. Features from the DNA-binding proteins can be extracted in a number of ways, such as, by protein sequence information, employing physicochemical properties of the amino acids (AA) and by using evolutionary information of the protein sequences. Sequence information and physicochemical properties are rapidly applied for protein function prediction [9,12–15]. However, recent studies have focused more on the use of evolutionary information and have obtained promising results [16–21].

To extract evolutionary information, profiles are computed from the local sequence alignment of the protein sequences. Some of the popular tools used for local sequence alignment are PSI-Blast and HHblits [22, 23]. These tools search a large database and build multiple sequence alignments (MSAs). From the MSAs, either the position-specific scoring matrix (PSSM) or the HMM profiles are generated, and the features are extracted. PSSM and HMM profiles are widely used in bioinformatics for the prediction of protein function [13,19,20,24–31]. Many studies have been reported to predict DNA-binding proteins [8,32–37] and DNA-binding protein sites [10,11,38–45]. However, only a few studies are reported to predict SSBs and DSBs [9,46,47]. Wang et al. [9], utilized the physicochemical properties and the PSSM profiles of the DNA-binding proteins to analyses the SSBs and DSBs. They extracted features including split amino acid transformation and dipeptide composition features from the protein sequences and applied support vector machine (SVM) and random forest (RF) classifiers to predict the SSBs and DSBs, respectively. In the recent study, Ali et al. [47], proposed SDBP-Pred predictor to predict the SSB and DSB proteins. They obtained the features from PSSM by applying the notion of consensus sequences and strategies of K-segmentation approach. The SDBP-Pred predictor achieved minor performance improvement on the benchmark data sets. Despite many efforts made to predict SSBs and DSBs accurately, the prediction performance reported is low. To enhance the performance of predicting DNA-binding proteins, machine learning methods such as the application of DeepInsight [48] method and the evaluation of evolutionary features of the DNA-binding protein sequences is required.

In this study, we utilize the HMM profiles generated using HHblits to compute the features of DNA-binding proteins. We utilized the normalized profile-monogram and normalized profile-bigram based feature extraction techniques [18] to compute the features. The profile-monogram and profile-bigram are well-known feature extraction techniques and has been extensively applied for protein fold recognition, subcellular localization, MoRF detection and protein drug target prediction [12–14,16–18,49]. For classification, SVM, k-nearest neighbors (KNN) and RF classifiers are used. These classifiers are widely used in bioinformatics and machine learning applications [50–53]. Two novel schemes are involved in the study, which makes a good predictive scheme for DNA-binding protein identification. First is the use of HMM profiles, which has not been explored for this study, and second is the use of normalized profile-monogram and normalized profile-bigram feature extraction methods which extracts the useful features encoded in the DNA-binding proteins. The proposed approach achieved promising results compared to the benchmarked method in the literature.

2. Method

2.1. Benchmark dataset

We used the training and independent sets that were previously introduced by Wang et al., [9]. The training set contains 1055 protein sequences, of which 183 are SSBs, and 873 are DSBs. To assemble this set, Wang et al. [9], collected a large number of DNA-binding proteins from UniProtKB, and Swiss-Prot databases, by manually reviewing the entries from the literature. Then, they used the CD-HIT tool [54] to extract the non-redundant proteins with sequence identity cut-off value of 0.7. The independent set contains 166 proteins, of which 41 are SSBs, and 125 are DSBs. They obtained this set from the protein data bank (PDB) and used the PISCES [55] tool to obtain non-redundant proteins with sequence similarity lower than 30%. The structure of the protein sequences in the independent set is determined experimentally by X-ray and NMR methods [9]. To obtain the protein sequences of the training and independent sets, we used the protein IDs from Wang et al. [9], to run the query search against the UniProt (www.uniprot.org) and PDB (www.rcsb.org) databases. The distribution of the SSBs and DSBs in the two datasets are shown in Fig. 1. To incorporate new SSB and DSB protein sequences, we collected 6988 SSBs and 6404 DSBs from UniProt database retrieved on July 20, 2020. The sequences are filtered by the search of the protein name as single-stranded binding and double-stranded binding, respectively. The search resulted with 249 SSBs and 58 DSBs, respectively. We, then used the CD-HIT tool [54] to compare the similarity of the sequences with the train and independent sets, the sequence identity cut-off value of 0.7 is used. To obtain the non-redundant proteins, the sequence similarity within the set is

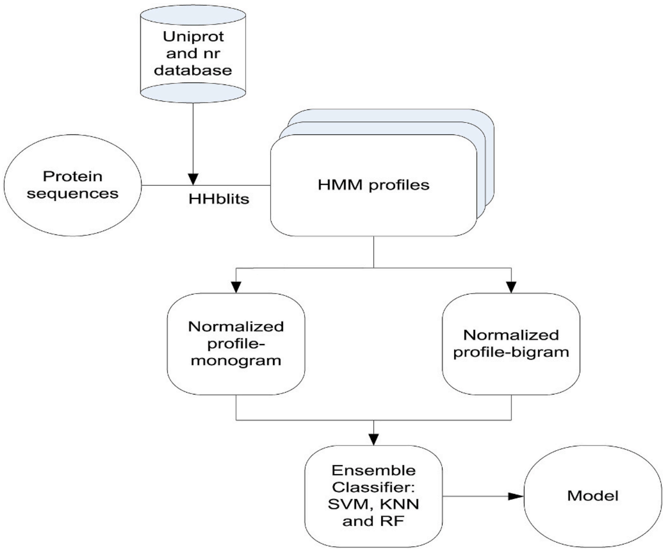
Uniprot and nr database
Protein sequences
HHblits
HMM profiles
Normalized profile-monogram
Normalized profile-bigram
Ensemble Classifier: SVM, KNN and RF
Model

Fig. 2. Overview of the proposed method.

checked with the sequence identity cut-off value of 0.9, 0.7 and 0.3, respectively, and for each the performance is reported.

2.2. Overview of the proposed method

Computational prediction of SSBs and DSBs requires the development of machine learning methods, which heavily depends on the feature extraction and classification algorithms. Features are extracted to represent the DNA-binding proteins. In classification, these features are used to predict the SSBs and DSBs. Features representing the DNA-binding protein sequence can be obtained in many ways, i.e., using sequence information, employing physicochemical properties of the amino acids, using structural information or evolutionary information of the protein sequence. For physicochemical properties of the amino acids, usually, the 544 physicochemical indexes are utilized to compute the physicochemical features [56]. On the other hand, for structural information, the attributes such as the output of Spider 2 (structural predictor) has been recently used [57]. The use of evolutionary information is gaining moment and has provided improved prediction accuracies [13,16,18,19]. The evolutionary information includes the PSSM profiles extracted by PSI-Blast tool [22] or the HMM profiles derived by HHblits tool [23].

In this study, we have utilized the HMM profiles to predict SSBs and DSBs. These profiles have not been explored for the employed dataset and have notably achieved good performance in related studies [13,19,58]. The HMM profiles of the protein sequences are used to derive normalized profile-monogram and normalized profile-bigram features. Classifiers, including SVM, KNN and RF, have been used for prediction. The overview of the proposed method is shown in Fig. 2. In the following sections, we describe the HMM profiles, the feature extraction methods, and the experimentation procedures.

2.2.1. HMM profiles

To generate the HMM profiles, HHblits [23] has been used. HHblits produces HMM profiles H of size L by 30, where L is the length of the protein sequence. HHblits is developed by Remmert et al. [23], and it iteratively searches through the databases and builds MSAs. From the MSAs, the HMM profiles are computed to represent the 20 standard amino acids in the homologous protein. Compared to the PSSM profiles, the HMM profiles contain additional information describing the insertion, deletion and match during MSAs. However, for this study, we only use the first 20 columns of the HMM profiles that represent the 20 common amino acids. To obtain the HMM profiles, we use the nr20 and uniprot20 databases, respectively, with HHblits cut-off value set to 0.001.

2.2.2. Feature extraction method

To compute features from the HMM profiles, we use the normalized profile-monogram and normalized profile-bigram [18] feature extraction methods. These methods are illustrated as follows:

- Normalized profile-monogram: using this method, the feature is computed from the HMM profiles of the protein sequences. Let matrix H of size $L \times 30$ be the HMM profile of a protein sequence. The computation of the feature vector from the matrix H is as follows:

$$NM(k) = \frac{1}{L}\sum_{i=1}^{L} H_{i,k}(1 \le k \le 20\ ) \tag{1}$$

where $H_{i,k}$ is the element of the HMM profile matrix. Computing $NM(k)$ for k ranging from 1 to 20 would give a feature vector of dimension 20.

- Normalized profile-bigram [18]: using this method, the feature is computed from the HMM profiles of the protein sequences. The computation of the feature vector from the matrix H is as follows:

$$NB_{k,l} = \frac{1}{\mathrm{L}}\sum_{\mathrm{i}=1}^{\mathrm{L}-1} H_{i,k}\ H_{i+1,l}(1 \le k \le 20 \ \text{ and } \ 1 \le l \le 20) \tag{2}$$

Computing $NB_{k,l}$ for $k = 1,\ 2, \ldots., 20$ and $l = 1,\ 2, \ldots., 20$ would give a matrix of size 20 × 20. This matrix can be represented in a vector form by reshaping the 20 × 20 matrix into a vector of length 400.

The use of normalized profile-monogram and normalized profile-bigram feature extraction methods have shown promising results for protein fold recognition, MoRF prediction, subcellular localization, drug-interaction and other related problems [12,16,18,49].

2.2.3. Experimentation

To show the effectiveness of the proposed method, we employ SVM, KNN and RF classifiers, respectively. These classifiers are widely used and have obtained good results in many related problems [14,16]. For experimentation, the LibSVM [59] package is adopted with radial basis function (RBF) kernel, and a grid search is performed to select the kernel parameters. For RF, the number of trees is set to 3000. For the prediction of the SSBs and DSBs, features from the DNA-binding proteins are extracted, and the classification models are trained.

To evaluate and measure the statistical significance of the proposed method, we adopted the 10-fold cross-validation method and performed experiments on the employed dataset. The training set is not balanced; thus, we use the random down-sampling technique to select an equal number of SSB and DSB samples. To report the performance, we repeated the 10-fold cross-validation method 50 times, each time randomly applying the down-sampling technique. The performance measures reported in this study include sensitivity (Sen), specificity (Spe), area under the ROC curve (AUC), Matthews correlation coefficients (MCC), accuracy (ACC) and F-measure (F1). The training set is used to evaluate the proposed method, while, the independent and test sets are used to provide an unbiased evaluation of the model.

To select the kernel parameters of the SVM classifier, we computed the amino acid composition (AAC) feature vector from the amino acids of the protein sequence. We used this feature vector to run a grid search using the LibSVM package [59], and the resulting C and gamma parameters of the kernel were found to be 4096 and 0.0029, respectively. To compare the performance of the proposed method with that of the recent method on the employed data sets, we implemented the algorithms of Wang et al. [9], and reported the performances. However, for the SDBP-pred predictor, the source code was not available, therefore, the performance is reported from Ali et al. [47], for the training and independent sets, respectively. To enhance the prediction of SSBs and

Table 1
The 10-fold cross-validation performance measures for different features on training dataset.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.6832	0.9409	0.9069	0.6464	0.8120	0.7838
	RF	0.8206	0.9148	0.9239	0.7389	0.8677	0.8611
Normalized profile-monogram nr20	SVM	0.8245	0.8732	0.8968	0.6988	0.8489	0.8452
	RF	0.8637	0.9033	0.959	0.7678	0.8835	0.8811
	KNN	**0.8815**	**0.8752**	**0.9494**	**0.7573**	**0.8784**	**0.8787**
	Ensemble method 1	0.7772	0.9650	0.9059	0.7558	0.8711	0.8577
	Ensemble method 2	0.8449	0.8975	0.9410	0.7437	0.8712	0.8678
Normalized profile-monogram uniprot20	SVM	0.8289	0.8596	0.8983	0.6889	0.8442	0.8418
	RF	0.8727	0.9071	0.9628	0.7804	0.8899	0.8880
	KNN	0.8833	0.8644	0.9516	0.7481	0.8738	0.8750
	Ensemble method 1	**0.7812**	**0.9584**	**0.9084**	**0.7516**	**0.8698**	**0.8571**
	Ensemble method 2	0.8622	0.8900	0.9446	0.7526	0.8761	0.8743
Normalized profile-bigram nr20	SVM	0.8141	0.8986	0.9052	0.7157	0.8563	0.8500
	RF	0.8546	0.9214	0.9685	0.7779	0.8880	0.8841
	KNN	0.8890	0.8467	0.9403	0.7372	0.8678	0.8706
	Ensemble method 1	0.7463	0.9712	0.8770	0.7365	0.8587	0.8407
	Ensemble method 2	0.8425	0.8984	0.9371	0.7424	0.8704	0.8667
Normalized profile-bigram uniprot20	SVM	**0.8235**	**0.8960**	**0.9107**	**0.7219**	**0.8597**	**0.8544**
	RF	**0.8539**	**0.9280**	**0.9701**	**0.7842**	**0.8909**	**0.8867**
	KNN	0.8978	0.8409	0.9429	0.7406	0.8693	0.8730
	Ensemble method 1	0.7588	0.9718	0.8892	0.7480	0.8653	0.8491
	Ensemble method 2	**0.8689**	**0.8984**	**0.9429**	**0.7680**	**0.8836**	**0.8819**

Bold numbers indicate the best performance for each classifier for different methods.

Table 2
Results for independent test.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.5020	0.7846	0.6874	0.2752	0.7148	0.4655
	RF	0.5200	0.7757	0.7120	0.2809	0.7125	0.4717
Normalized profile-monogram nr20	SVM	0.6678	0.744	0.7492	0.3723	0.7252	0.5461
	RF	0.5629	0.7602	0.7518	0.3003	0.7115	0.4916
	KNN	**0.6795**	**0.6706**	**0.7444**	**0.3080**	**0.6728**	**0.5066**
	Ensemble method 1	**0.4463**	**0.8928**	**0.6887**	**0.3752**	**0.7825**	**0.504**
	Ensemble method 2	**0.6405**	**0.737**	**0.7588**	**0.3408**	**0.7131**	**0.5245**
Normalized profile-monogram uniprot20	SVM	0.6981	0.7008	0.7438	0.3529	0.7001	0.5358
	RF	0.5395	0.7602	0.7410	0.2803	0.7057	0.4754
	KNN	0.6127	0.6768	0.7106	0.2582	0.6610	0.473
	Ensemble method 1	0.4576	0.8891	0.7058	0.3795	0.7825	0.5099
	Ensemble method 2	0.6107	0.7149	0.7254	0.2926	0.6892	0.4928
Normalized profile-bigram nr20	SVM	**0.5976**	**0.7925**	**0.7577**	**0.3689**	**0.7443**	**0.5356**
	RF	0.5971	0.7720	0.7694	0.3424	0.7288	0.5205
	KNN	0.6639	0.5933	0.7056	0.2246	0.6107	0.4591
	Ensemble method 1	0.4093	0.8933	0.6377	0.3415	0.7737	0.4715
	Ensemble method 2	0.6146	0.7211	0.7285	0.3031	0.6948	0.4986
Normalized profile-bigram uniprot20	SVM	0.6010	0.7805	0.7594	0.3577	0.7361	0.5298
	RF	**0.6054**	**0.7709**	**0.7697**	**0.3484**	**0.7300**	**0.5254**
	KNN	0.6985	0.5678	0.7209	0.2316	0.6001	0.4652
	Ensemble method 1	0.4327	0.8805	0.6526	0.3429	0.7699	0.4814
	Ensemble method 2	0.6624	0.7016	0.7528	0.3244	0.6919	0.5158

Bold numbers indicate the best performance for each classifier for different methods.

DSBs, the ensemble of the classifier is developed. The first ensemble is developed by using the concept of majority voting. SVM, RF and KNN classifiers have been used in this work. Therefore, the majority of these three classifiers classes is used to determine the final class for a test sample. The second ensemble is developed by looking at the class probabilities of the three classifiers. Out of the three classifiers class probabilities, the classifier giving the highest class probability of a test sample is used to determine the final class of the test sample.

3. Results and discussion

To predict the SSB and DSB proteins, recent studies utilized the physicochemical properties and the PSSM profiles of the DNA-binding proteins. Wang et al. [9], extracted features including split amino acid transformation and dipeptide composition features from the protein sequences and applied SVM and RF classifiers for prediction. Ali et al. [47], proposed SDBP-Pred predictor to predict the SSB and DSB proteins. They obtained features from PSSM by applying the notion of consensus sequences and strategies of K-segmentation approach. The SDBP-Pred predictor presented small improvement on the performance compared with the Wang et al. [9], method. To further enhance the performance, in this study, HMM profiles of the protein sequence is used for feature extraction. Normalized profile-monogram and normalized profile-bigram based features are extracted and the classifiers including SVM, KNN and RF are used for prediction. Tables 1 and 2 show the performance measures for the various methods adopted in this study. In Table 1, the result is reported for performing 10-fold cross-validation method on the training set and in Table 2, the result is shown for the independent test. The results are reported for normalized profile-monogram and normalized profile-bigram features computed from the HMM profiles of the protein sequences. Comparing the results with the recent benchmark method [9], a consistent increase in the performance is observed for the proposed method.

Evaluating the proposed method on the training set, the normalized

Table 3
Comparison with existing methods on train set.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.6832	0.9409	0.9069	0.6464	0.812	0.7838
	RF	0.8206	0.9148	0.9239	0.7389	0.8677	0.8611
SDBP-Pred (2020)	SVM	0.9427	0.8033	–	0.7220	0.9186	–
Proposed	SVM	0.8235	0.8960	0.9107	0.7219	0.8597	0.8544
	RF	0.8539	0.9280	0.9701	0.7842	0.8909	0.8867
	KNN	0.8815	0.8752	0.9494	0.7573	0.8784	0.8787
	Ensemble method 1	0.7812	0.9584	0.9084	0.7516	0.8698	0.8571
	Ensemble method 2	0.8689	0.8984	0.9429	0.7680	0.8836	0.8819

The underlined scores are obtained from the Ali et al. [47], since the source code for SDBP-Pred predictor is not available.

Table 4
Comparison with existing methods on independent test set.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.502	0.7846	0.6874	0.2752	0.7148	0.4655
	RF	0.5200	0.7757	0.7120	0.2809	0.7125	0.4717
SDBP-Pred (2020)	SVM	0.8160	0.4870	–	0.2990	0.7340	–
Proposed	SVM	0.5976	0.7925	0.7577	0.3689	0.7443	0.5356
	RF	0.6054	0.7709	0.7697	0.3484	0.7300	0.5254
	KNN	0.6795	0.6706	0.7444	0.3080	0.6728	0.5066
	Ensemble method 1	0.4463	0.8928	0.6887	0.3752	0.7825	0.5040
	Ensemble method 2	0.6405	0.7370	0.7588	0.3408	0.7131	0.5245

The underlined scores are obtained from the Ali et al. [47], since the source code for SDBP-Pred predictor is not available.

profile-bigram features computed from the HMM profiles (extracted from uniprot database) performed well with the SVM and RF classifiers, respectively, achieving a sensitivity of 0.824 and 0.854, specificity of 0.896 and 0.928, AUC of 0.911 and 0.970, MCC of 0.722 and 0.784, accuracy of 0.860 and 0.891, and F-measure of 0.854 and 0.887. On the other hand, normalized profile-monogram features computed from the HMM profiles of the nr20 database performed well with the KNN classifier achieving a sensitivity of 0.882, specificity of 0.875, AUC of 0.949, MCC of 0.757, accuracy of 0.878, and F-measure of 0.879. Overall, a performance improvement of 5–10% is observed compared with the method reported by Wang et al., [9]. In the independent test, the normalized profile-bigram features computed from the HMM profiles of the nr20 and uniprot databases provided good results with the SVM and RF classifiers, respectively, achieving a sensitivity of 0.600 and 0.605, specificity of 0.793 and 0.771, AUC of 0.758 and 0.769, MCC of 0.369 and 0.348, accuracy of 0.744 and 0.730, and F-measure of 0.56 and 0.525. Similarly, the normalized profile-monogram features computed from the HMM profiles of the nr20 database performed well with the KNN classifier achieving a sensitivity of 0.680, specificity of 0.671, AUC of 0.744, MCC of 0.308, accuracy of 0.673, and F-measure of 0.507. Compared to the benchmark method proposed by Wang et al. [9], our proposed method demonstrated a performance improvement of approximately more than 3%.

To enhance the overall performance, the ensemble of the classifiers is developed, and a minor increase in performance is reported. Evaluating the train set, the ensemble method 1 performed well with normalized profile-monogram feature achieving sensitivity of 0.781, specificity of 0.958, AUC of 0.908, MCC of 0.752, accuracy of 0.870, and F-measure of 0.857. On the other hand, the ensemble method 2 provided good results with the normalized profile-bigram feature achieving a sensitivity of

Table 5
Results for new test set for sequences identity cut off at 90%.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.7653	0.9306	0.856	0.6970	0.8402	0.8399
	RF	0.7641	0.9491	0.9127	0.7155	0.8479	0.8462
Normalized profile-monogram nr20	SVM	0.7522	0.8068	0.8371	0.5579	0.7769	0.7871
	RF	**0.8456**	**0.9257**	**0.8912**	**0.7684**	**0.8819**	**0.8867**
	KNN	**0.8241**	**0.7589**	**0.8853**	**0.5861**	**0.7945**	**0.8149**
	Ensemble method 1	**0.7341**	**0.9947**	**0.8228**	**0.7396**	**0.8521**	**0.8445**
	Ensemble method 2	**0.8109**	**0.8740**	**0.8677**	**0.6833**	**0.8395**	**0.8472**
Normalized profile-monogram uniprot20	SVM	0.7500	0.7921	0.8366	0.5407	0.7691	0.7810
	RF	0.8409	0.8615	0.8803	0.7010	0.8503	0.8598
	KNN	0.8272	0.6109	0.8788	0.4521	0.7292	0.7699
	Ensemble method 1	0.7294	0.9785	0.8391	0.7171	0.8422	0.8350
	Ensemble method 2	0.8088	0.7717	0.8681	0.5817	0.7920	0.8098
Normalized profile-bigram nr20	SVM	**0.7409**	**0.8909**	**0.8305**	**0.6326**	**0.8089**	**0.8096**
	RF	0.7697	0.8034	0.8422	0.5718	0.7850	0.7971
	KNN	0.8038	0.5132	0.8365	0.3346	0.6721	0.7286
	Ensemble method 1	0.7288	0.9377	0.8531	0.6715	0.8234	0.8189
	Ensemble method 2	0.7556	0.7306	0.8296	0.4863	0.7443	0.7644
Normalized profile-bigram uniprot20	SVM	0.7338	0.8853	0.8254	0.620	0.8024	0.8028
	RF	0.7713	0.7868	0.8329	0.5570	0.7783	0.7926
	KNN	0.8125	0.4800	0.8314	0.3124	0.6619	0.7247
	Ensemble method 1	0.7281	0.9340	0.8425	0.6668	0.8214	0.8170
	Ensemble method 2	0.7775	0.6925	0.8288	0.4726	0.7390	0.7658

Bold numbers indicate the best performance for each classifier for different methods.

Table 6
Results for new test set for sequences identity cut off at 70%.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.2967	0.9119	0.4969	0.2718	0.7612	0.3810
	RF	0.2400	0.9454	0.6114	0.2794	0.7727	0.3417
Normalized profile-monogram nr20	SVM	0.2433	0.7897	0.5239	0.0427	0.6559	0.2573
	RF	0.4933	0.9238	0.6839	0.4679	0.8184	0.5672
	KNN	**0.5150**	**0.7432**	**0.6946**	**0.2451**	**0.6874**	**0.4474**
	Ensemble method 1	**0.1800**	**0.9968**	**0.4801**	**0.3611**	**0.7967**	**0.2967**
	Ensemble method 2	**0.4750**	**0.8876**	**0.6373**	**0.4014**	**0.7865**	**0.5249**
Normalized profile-monogram uniprot20	SVM	**0.3133**	**0.7714**	**0.5394**	**0.0942**	**0.6592**	**0.3156**
	RF	**0.5333**	**0.8935**	**0.6565**	**0.4501**	**0.8053**	**0.5692**
	KNN	0.5250	0.5741	0.6385	0.0864	0.5620	0.3691
	Ensemble method 1	0.2117	0.9897	0.5336	0.3711	0.7992	0.3381
	Ensemble method 2	0.420	0.7773	0.6164	0.1973	0.6898	0.3991
Normalized profile-bigram nr20	SVM	0.2083	0.9022	0.4646	0.1548	0.7322	0.2783
	RF	0.1767	0.7935	0.4511	–	0.6425	0.1953
	KNN	0.3450	0.4881	0.4462	–	0.4531	0.2253
	Ensemble method 1	0.1433	0.9600	0.6101	0.1923	0.760	0.2249
	Ensemble method 2	0.1950	0.7108	0.4432	–	0.5845	0.1874
Normalized profile-bigram uniprot20	SVM	0.200	0.9097	0.4606	0.1566	0.7359	0.2706
	RF	0.1933	0.7719	0.4137	–	0.6302	0.2091
	KNN	0.4067	0.4254	0.4312	–	0.4208	0.2511
	Ensemble method 1	0.1700	0.9687	0.6046	0.2589	0.7731	0.2684
	Ensemble method 2	0.2417	0.6654	0.4176	–	0.5616	0.2115

Bold numbers indicate the best performance for each classifier for different methods.

Table 7
Results for new test set for sequences identity cut off at 30%.

Method	Classifier	Sen	Spec	AUC	MCC	ACC	F1
Wang et al. (BMC 2017)	SVM	0.3933	0.9400	0.6102	0.4303	0.776	0.5143
	RF	0.3200	0.9629	0.6801	0.4067	0.7700	0.4548
Normalized profile-monogram nr20	SVM	0.3044	0.8743	0.6265	0.2205	0.7033	0.3807
	RF	0.5267	0.8733	0.6965	0.4252	0.7693	0.5748
	KNN	**0.5289**	**0.7781**	**0.7396**	**0.3077**	**0.7033**	**0.5141**
	Ensemble method 1	0.2267	0.9952	0.5297	0.3949	0.7647	0.3592
	Ensemble method 2	**0.5089**	**0.8752**	**0.6907**	**0.4183**	**0.7653**	**0.5660**
Normalized profile-monogram uniprot20	SVM	**0.4178**	**0.8591**	**0.6526**	**0.3123**	**0.7267**	**0.4805**
	RF	**0.6044**	**0.8667**	**0.694**	**0.4862**	**0.7880**	**0.6295**
	KNN	0.5956	0.6857	0.7353	0.2652	0.6587	0.5110
	Ensemble method 1	**0.2822**	**0.9857**	**0.5328**	**0.4229**	**0.7747**	**0.4253**
	Ensemble method 2	0.5000	0.7771	0.6986	0.2785	0.6940	0.4944
Normalized profile-bigram nr20	SVM	0.2778	0.9219	0.5510	0.2771	0.7287	0.3821
	RF	0.2356	0.7457	0.4933	–	0.5927	0.2556
	KNN	0.4044	0.5191	0.5104	–	0.4847	0.309
	Ensemble method 1	0.1911	0.9400	0.6360	0.2084	0.7153	0.2842
	Ensemble method 2	0.2578	0.6981	0.5127	–	0.5660	0.2616
Normalized profile-bigram uniprot20	SVM	0.2667	0.9152	0.5597	0.2455	0.7207	0.3622
	RF	0.2556	0.7476	0.4618	0.006	0.60	0.2774
	KNN	0.4422	0.5114	0.5199	–	0.4907	0.3363
	Ensemble method 1	0.2267	0.9591	0.6617	0.2957	0.7393	0.3406
	Ensemble method 2	0.3222	0.6933	0.5114	0.0147	0.5820	0.3124

Bold numbers indicate the best performance for each classifier for different methods.

0.869, specificity of 0.898, AUC of 0.943, MCC of 0.768, accuracy of 0.884, and F-measure of 0.882. In the independent set, both the ensemble methods performed well with the normalized profile-monogram features. To measure the statistical significance of the proposed method, we performed paired *t*-test with 5% significance level. For the 10-fold cross-validation and independent test experiments, respectively, the statistical significance of the proposed method compared with Wang et al. [9] method, is computed as 0.028 and 0.008 for SVM classification, and 0.005 and 0.010 for RF classification. The performance measures including AUC, MCC, ACC and F-measure have been utilized to compute the statistical significance. The overall comparison of the result is shown in Table 3 and Table 4. The proposed method is compared with Wang et al. [9], method and the recent predictor SDBP-Pred [47]. Overall, it is noticed that the proposed method outperforms the benchmark method and achieves promising results. To assess the performance of the proposed method, a new test set is assembled to include the recent DNA-binding proteins. For the new assembled test set, the performance is reported in Tables 5–7, respectively. The sequence similarity cut-off value of 0.9, 0.7 and 0.3 is utilized to report the performance. It is observed that the proposed method outperforms the Wang et al. [9], method with a good prediction performance. The sources code for SDBP-pred predictor is not available publicly, therefore, its performance is not compared for the new test set.

The results suggest that the essential properties of SSBs and DSBs can be revealed by the profile monogram and profile bigram features extracted from the HMM profiles of the protein sequences. This helped in achieving improved performance for the prediction of the SSBs and DSBs. These properties are successfully encoded in the predicting scheme using the feature extraction techniques adopted in this study. To visualize these properties, we analyze the normalized profile-monogram and normalized profile-bigram scores computed from the HMM profiles of the protein sequences (Fig. 3 and Fig. 4). These scores are computed for the SSBs and DSBs in the independent set. The normalized profile-monogram features clearly demonstrate the difference in SSBs and

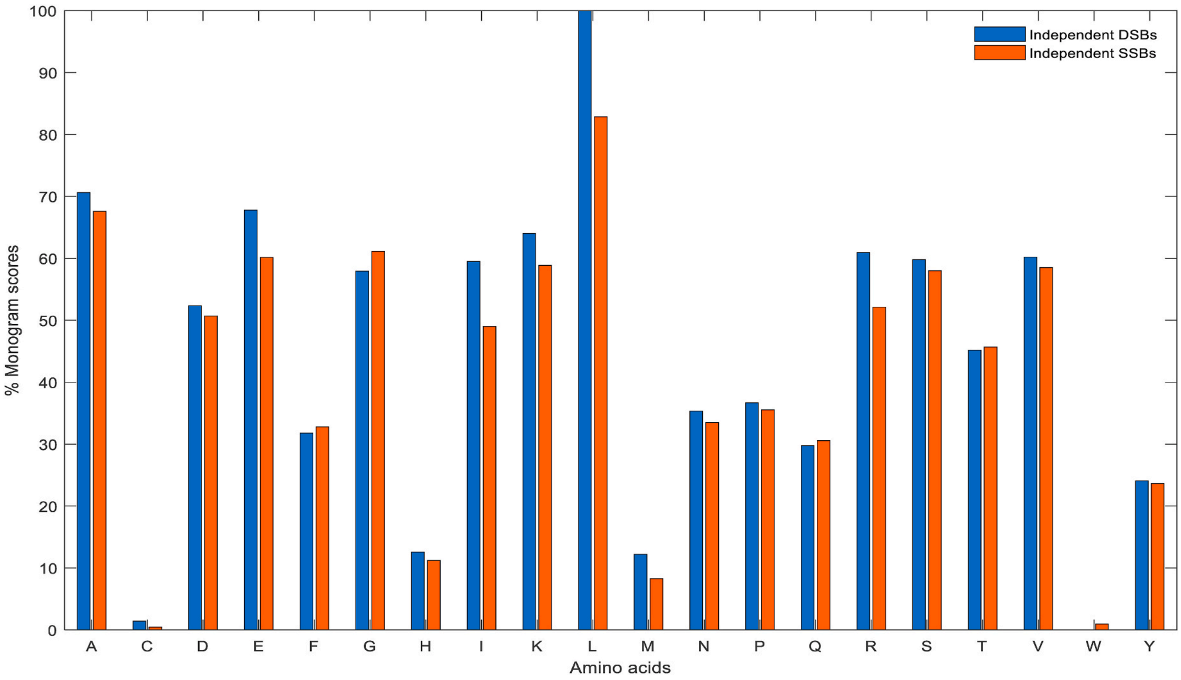

Fig. 3. Normalized profile-monogram scores of the SSBs and DSBs in the independent dataset. The average value of the scores are shown.

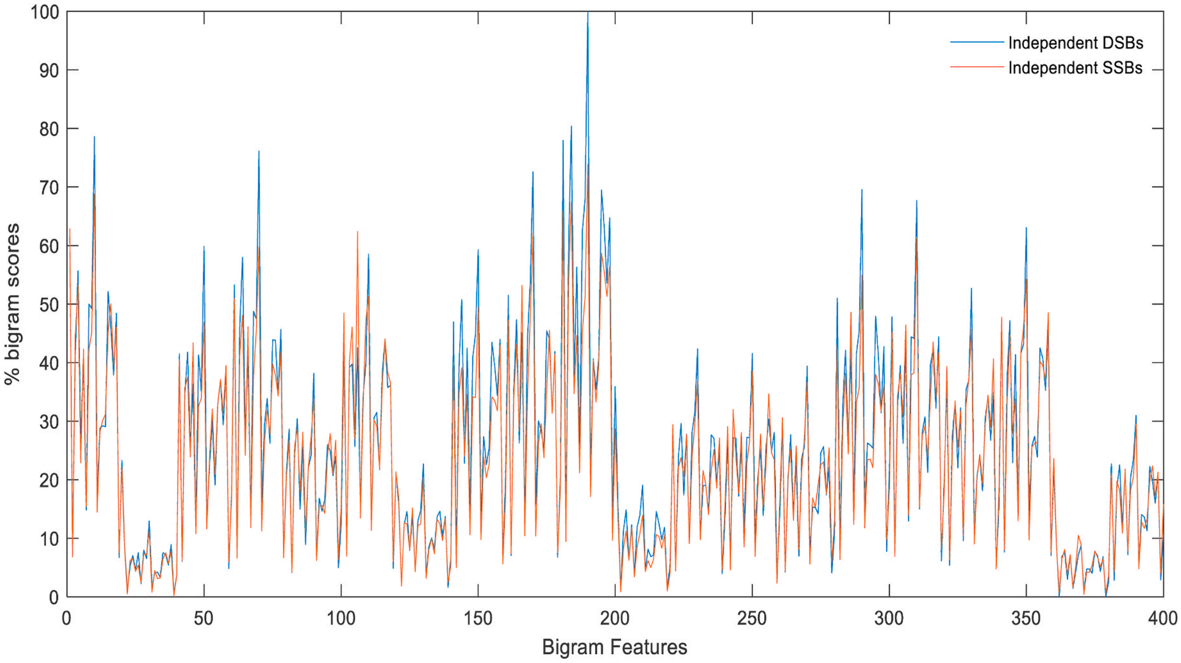

Fig. 4. Normalized profile-bigram scores of the SSBs and DSBs in the independent dataset. The average value of the scores are shown.

DSBs for the amino acids E, I, L, K and R, while the amino acids A, G, M, N and S show a close relationship. Furthermore, we observe that amino acids D, F, H, P, Q, T, V and Y have similar properties in SSBs and DSBs. Besides, the normalized profile-bigram features (Fig. 4) demonstrate the difference in the properties of SSBs and DSBs, which helped the proposed method to predict the SSBs and DSBs correctly.

HMM profiles generated with HHblits contain relevant information used to predict the SSBs and DSBs. This is shown computationally by observing the HMM profile features compared with the features of the primary protein sequences (Fig. 5). In Fig. 5, it is observed that the SSBs and DSBs are easily recognized with the HMM profile features, whereas, it is difficult to identify the SSBs and DSBs using the features of the primary protein sequences. HMM profiles contain the evolutionary information of protein sequences in the external databases and its significance lies in the construction of high-quality multiple sequences alignments. Comparing the prediction performance, the normalized profile-bigram based features dominated the accuracy with different classifiers and parameters, therefore, the normalized profile-bigram

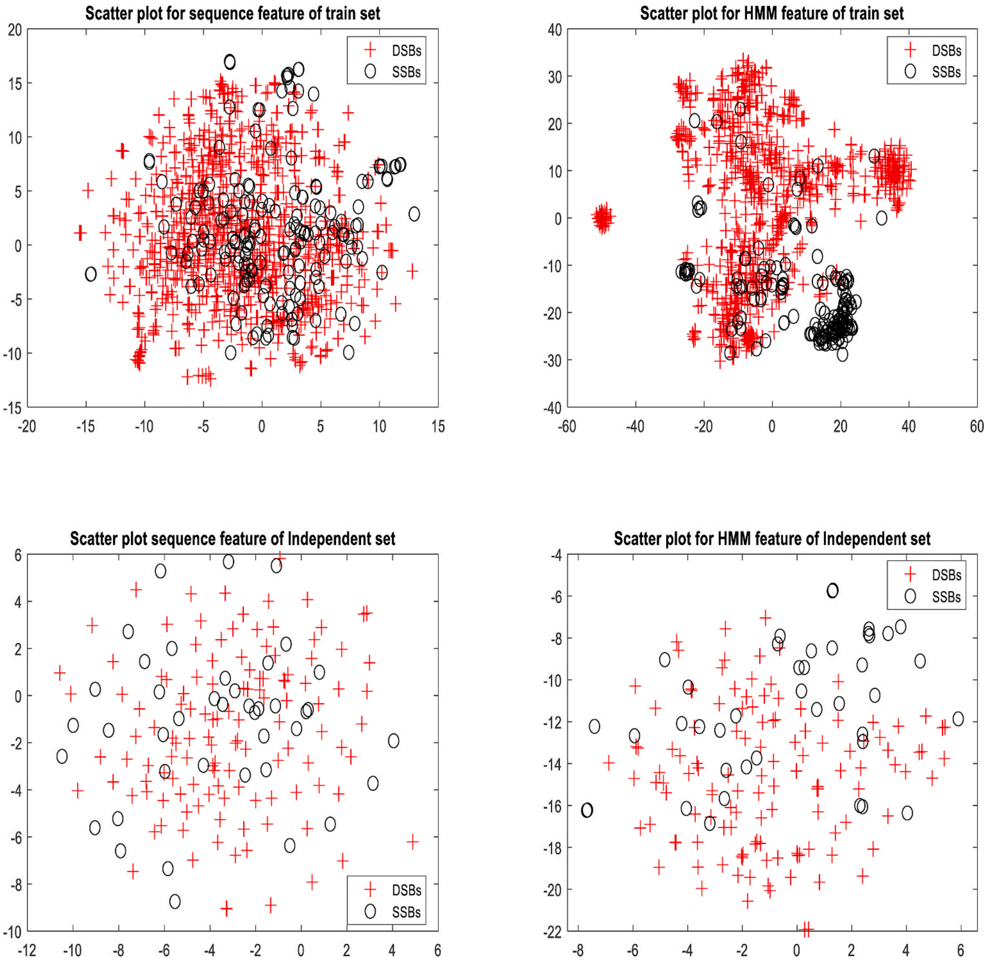

Fig. 5. Two-dimensional plots of the features of SSBs and DSBs in the train and independent sets. Tsne algorithm is used to reduce the feature dimension to obtain the two-dimensional plot.

based feature extraction method is recommended for the proposed method to predict SSBs and DSBs.

We are currently investigating the AI techniques, such as DeepInsight [48], to predict the DNA-binding proteins. Similar to our previous studies [13,16,17,19], in future, we will be identifying and analyzing the amino acid residues involved in DNA-binding sites. We are aiming to develop a computational predictor to predict the DNA-binding sites.

4. Conclusion

In this study, the HMM profiles are used for the prediction of SSBs and DSBs. The comparison of the results demonstrates the significance of the proposed method. The results revealed the distinguishing abilities of the profile-monogram and profile-bigram features computed from the HMM profiles. The features showed a remarkable difference between SSBs and DSBs. Using these features and conducting an independent test, confirmed the effectiveness of the proposed method. Compared with the RF and KNN classifier employed in this study, SVM classifier performed better in the independent test. The proposed method achieved a performance improvement of approximately 3% in the independent test compared with the existing method [9], thus, indicating the effective prediction of SSBs and DSBs to investigate the DNA-binding proteins.

Funding

This article is funded by RIKEN, Center for Integrative Medical Sciences, Japan and CREST, JST, Yokohama 230–0045, Japan.

Authors' contributions

RS, SK, TK and AS conceived the project. RS processed the data. RS, SK and AS performed the analysis and wrote the manuscript. TT provided computational resources. All authors read and approved the final manuscript.

Declaration of competing interest

We have no competing interest.

References

[1] N.M. Luscombe, S.E. Austin, H.M. Berman, J.M. Thornton, An overview of the structures of protein-DNA complexes, Genome Biol. 1 (1) (2000) reviews001.001.

[2] J. Rhodin Edsö, C. Gustafsson, M. Cohn, Single- and double-stranded DNA binding proteins act in concert to conserve a telomeric DNA core sequence, Genome Integr. 2 (1) (2011), 2-2.

[3] L. Attaiech, A. Olivier, I. Mortier-Barriere, A.L. Soulet, C. Granadel, B. Martin, P. Polard, J.P. Claverys, Role of the single-stranded DNA-binding protein SsbB in pneumococcal transformation: maintenance of a reservoir for genetic plasticity, PLoS Genet. 7 (6) (2011) 30.

[4] L.S. Shlyakhtenko, A.Y. Lushnikov, A. Miyagi, Y.L. Lyubchenko, Specificity of binding of single-stranded DNA-binding protein to its target, Biochemistry 51 (7) (2012) 1500–1509.

[5] D.J. Richard, E. Bolderson, L. Cubeddu, R.I. Wadsworth, K. Savage, G.G. Sharma, M.L. Nicolette, S. Tsvetanov, M.J. McIlwraith, R.K. Pandita, et al., Single-stranded DNA-binding protein hSSB1 is critical for genomic stability, Nature 453 (7195) (2008) 677–681.

[6] Y. Ofran, V. Mysore, B. Rost, Prediction of DNA-binding residues from sequence, Bioinformatics 23 (13) (2007) i347–i353.

[7] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.

[8] B. Liu, S. Wang, X. Wang, DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation, Sci. Rep. 5 (15479) (2015).

[9] W. Wang, L. Sun, S. Zhang, H. Zhang, J. Shi, T. Xu, K. Li, Analysis and prediction of single-stranded and double-stranded DNA binding proteins based on protein sequences, BMC Bioinf. 18 (1) (2017), 017-1715.

[10] Q. Zhou, J.S. Liu, Extracting sequence features to predict protein-DNA interactions: a comparative study, Nucleic Acids Res. 36 (12) (2008) 4137–4148.

[11] W. Xiong, T. Li, K. Chen, K. Tang, Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information, Nucleic Acids Res. 37 (17) (2009) 5632–5640.

[12] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda, A. Sharma, Predict Gram-positive and Gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into Chou's general PseAAC, IEEE Trans. NanoBioscience 14 (8) (2015) 915–926.

[13] R. Sharma, M. Bayarjargal, T. Tsunoda, A. Patil, A. Sharma, MoRFPred-plus: computational identification of MoRFs in protein sequences using physicochemical properties and HMM profiles, J. Theor. Biol. 437 (Supplement C) (2018) 9–16.

[14] A. Dehzangi, S. Sohrabi, R. Heffernan, A. Sharma, J. Lyons, K.K. Paliwal, A. Sattar, Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features, BMC Bioinf. 16 (4) (2014).

[15] S. Wan, M.W. Mak, S.Y. Kung, Ensemble linear neighborhood propagation for predicting subchloroplast localization of multi-location proteins, J. Proteome Res. 15 (12) (2016) 4755–4762.

[16] R. Sharma, G. Raicar, T. Tsunoda, A. Patil, A. Sharma, OPAL: prediction of MoRF regions in intrinsically disordered protein sequences, Bioinformatics 34 (11) (2018) 1850–1858.

[17] R. Sharma, A. Sharma, G. Raicar, T. Tsunoda, A. Patil, OPAL+: length-specific MoRF prediction in intrinsically disordered protein sequences, Proteomics (2018), https://doi.org/10.1002/pmic.201800058.

[18] A. Sharma, J. Lyons, A. Dehzangi, K.K. Paliwai, A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition, Theor. Biol. 320 (7) (2013) 41–46.

[19] R. Sharma, S. Kumar, T. Tsunoda, A. Patil, A. Sharma, Predicting MoRFs in protein sequences using HMM profiles, BMC Bioinf. 17 (Suppl X) (2016) S14.

[20] J. Lyons, A. Dehzangi, R. Heffernan, Y. Yang, Y. Zhou, A. Sharma, K. Paliwal, Advancing the accuracy of protein fold recognition by utilizing profiles from hidden Markov models, IEEE Trans. NanoBioscience 14 (7) (2015) 761–772.

[21] R. Sharma, A. Sharma, A. Patil, T. Tsunoda, Discovering MoRFs by trisecting intrinsically disordered protein sequence into terminals and middle regions, BMC Bioinf. 19 (13) (2019) 378.

[22] S.F. Altschul, T.L. Madden, A.A. Schaffer, J.H. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 17 (1997) 3389–3402.

[23] M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat. Methods 9 (2) (2011) 173–175.

[24] M. Delorenzi, T. Speed, An HMM model for coiled-coil domains and a comparison with PSSM-based predictions, Bioinformatics 18 (4) (2002) 617–625.

[25] K.C. Chou, H.B. Shen, MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochem. Biophys. Res. Commun. 360 (2) (2007) 339–345.

[26] S. Wan, M. Mak, S. Kung, Transductive learning for multi-label protein subchloroplast localization prediction, IEEE ACM Trans. Comput. Biol. Bioinf 14 (1) (2017) 212–224.

[27] A. Dehzangi, K.K. Paliwal, J. Lyons, A. Sharma, A. Scattar, A segmentation-based method to extract structural and evolutionary features for protein fold recognition, IEEE ACM Trans. Comput. Biol. Bioinf 11 (3) (2013) 510–519.

[28] H. Saini, G. Raicar, A. Sharma, S. Lal, A. Dehzangi, A. Rajeshkannan, J. Lyons, N. Biswas, K.K. Paliwal, Protein structural class prediction via k-separated bigrams using position specific scoring matrix, J. Adv. Comput. Intell. Intell. Inf. 18 (4) (2014) 474–479.

[29] A. Dehzangi, R. Hefterman, A. Sharma, J. Lyons, K.K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, Theor. Biol. 364 (2015) 284–294.

[30] M.W. Ahmad, M.E. Arafat, G. Taherzadeh, A. Sharma, S.R. Dipta, A. Dehzangi, S. Shatabda, Mal-light: enhancing lysine malonylation sites prediction problem using evolutionary-based features, IEEE Access 8 (2020) 77888–77902.

[31] A.A. Chandra, A. Sharma, A. Dehzangi, T. Tsunoda, EvolStruct-Phogly: incorporating structural properties and evolutionary information from profile bigrams for the phosphoglycerylation prediction, BMC Genom. 19 (9) (2019) 984.

[32] W. Zhou, H. Yan, Prediction of DNA-binding protein based on statistical and geometric features and support vector machines, Proteome Sci. 9 (Suppl 1) (2011) (Suppl 1):S1-S1.

[33] A. Szabóová, O. Kuželka, F. Železný, J. Tolar, Prediction of DNA-binding propensity of proteins by the ball-histogram method using automatic template search, BMC Bioinf. 13 (10) (2012) S3.

[34] G. Nimrod, A. Szilagyi, C. Leslie, N. Ben-Tal, Identification of DNA-binding proteins using structural, electrostatic and evolutionary features, J. Mol. Biol. 387 (4) (2009) 1040–1053.

[35] W.-Z. Lin, J.-A. Fang, X. Xiao, K.-C. Chou, iDNA-Prot, Identification of DNA binding proteins using random forest with grey model, PloS One 6 (9) (2011) e24756-e24756.

[36] S.Y. Chowdhury, S. Shatabda, A. Dehzangi, iDNAProt-ES: identification of DNA-binding proteins using evolutionary and structural features, Sci. Rep. 7 (1) (2017) 14938.

[37] M. Kumar, M.M. Gromiha, G.P.S. Raghava, Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioinf. 8 (2007), 463-463.

[38] I.B. Kuznetsov, Z. Gou, R. Li, S. Hwang, Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins, Proteins 64 (1) (2006) 19–27.

[39] J. Si, R. Zhao, R. Wu, An overview of the prediction of protein DNA-binding sites, Int. J. Mol. Sci. 16 (3) (2015) 5194–5215.

[40] Y. Xiong, J. Xia, W. Zhang, J. Liu, Exploiting a reduced set of weighted average features to improve prediction of DNA-binding residues from 3D structures, PloS One 6 (12) (2011), e28440.

[41] X. Zhu, S.S. Ericksen, J.C. Mitchell, DBSI: DNA-binding site identifier, Nucleic Acids Res. 41 (16) (2013) e160-e160.

[42] S. Dey, A. Pal, M. Guharoy, S. Sonavane, P. Chakrabarti, Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters, Nucleic Acids Res. 40 (15) (2012) 7150–7161.

[43] C. Yan, M. Terribilini, F. Wu, R.L. Jernigan, D. Dobbs, V. Honavar, Predicting DNA-binding sites of proteins from amino acid sequence, BMC Bioinf. 7 (1) (2006) 262.

[44] Y. Xiong, J. Liu, D.Q. Wei, An accurate feature-based method for identifying DNA-binding residues on protein surfaces, Proteins 79 (2) (2011) 509–517.

[45] Y. Cai, J. He, X. Li, L. Lu, X. Yang, K. Feng, W. Lu, X. Kong, A novel computational approach to predict transcription factor DNA binding preference, J. Proteome Res. 8 (2) (2009) 999–1003.

[46] W. Wang, J. Liu, X. Zhou, Identification of single-stranded and double-stranded dna binding proteins based on protein structure, BMC Bioinf. 15 (12) (2014) S4.

[47] F. Ali, M. Arif, Z.U. Khan, M. Kabir, S. Ahmed, D.J. Yu, SDBP-Pred: prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM, Anal. Biochem. 589 (113494) (2020) 3.

[48] A. Sharma, E. Vans, D. Shigemizu, K.A. Boroevich, T. Tsunoda, DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture, Sci. Rep. 9 (1) (2019) 11399.

[49] Z. Mousavian, S. Khakabimamaghani, K. Kavousi, A. Masoudi-Nejad, Drug–target interaction prediction from PSSM based evolutionary information, J. Pharmacol. Toxicol. Methods 78 (2016) 42–51.

[50] T. Ebina, H. Toh, Y. Kuroda, DROP: an SVM domain linker predictor trained with optimal features selected by random forest, Bioinformatics 27 (4) (2011) 487–494.

[51] S. Wan, M.W. Mak, S.Y. Kung, Mem-ADSVM, A two-layer multi-label predictor for identifying multi-functional types of membrane proteins, J. Theor. Biol. 398 (2016) 32–42.

[52] R. Jiang, W. Tang, X. Wu, W. Fu, A random forest approach to the detection of epistatic interactions in case-control studies, BMC Bioinf. 10 (Suppl 1) (2009) (Suppl 1):S65-S65.

[53] S. Wan, M.-W. Mak, Predicting subcellular localization of multi-location proteins by improving support vector machines with an adaptive-decision scheme, Int. J. Machine Learn. Cybern. 9 (3) (2018) 399–411.

[54] W. Li, A. Godzik, Cd-hit, A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (13) (2006) 1658–1659.

[55] G. Wang, R.L. Dunbrack Jr., PISCES: recent improvements to a PDB sequence culling server, Nucleic Acids Res. 33 (2005). Web Server issue.

[56] A. Sharma, K.K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, S. Miyano, A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition, BMC Bioinf. 14 (233) (2013) 1–11.

[57] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Zhou, SPIDER2: a package to predict secondary structure, accessible surface area and main-chain torsional angles by deep neural networks, Methods Mol. Biol. 1484 (2017) 55–63.

[58] J. Lyons, K.K. Paliwal, A. Dehzangi, R. Heffernan, T. Tsunoda, A. Sharma, Protein fold recognition using HMM-HMM alignment and dynamic programming, J. Theor. Biol. 393 (2016) 67–74.

[59] C.C. Chang, C.J. Lin, Libsvm : a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (27) (2011) 1–27.