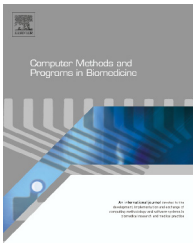
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 117 (2014) 158–167

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Sequence-based prediction of protein-binding sites in DNA: Comparative study of two SVM models

Byungkyu Park[a], *Jinyong Im*[b], *Narankhuu Tuvshinjargal*[b], *Wook Lee*[b], *Kyungsook Han*[b,*]

[a] *Institute for Information and Electronics Research, Inha University, Incheon, South Korea*
[b] *Department of Computer Science and Engineering, Inha University, Incheon, South Korea*

ARTICLE INFO

Article history:
Received 7 December 2013
Received in revised form
17 July 2014
Accepted 18 July 2014

Keywords:
DNA–protein interactions
Binding sites
Protein-binding nucleotides
Prediction model

ABSTRACT

As many structures of protein–DNA complexes have been known in the past years, several computational methods have been developed to predict DNA-binding sites in proteins. However, its inverse problem (i.e., predicting protein-binding sites in DNA) has received much less attention. One of the reasons is that the differences between the interaction propensities of nucleotides are much smaller than those between amino acids. Another reason is that DNA exhibits less diverse sequence patterns than protein. Therefore, predicting protein-binding DNA nucleotides is much harder than predicting DNA-binding amino acids. We computed the interaction propensity (IP) of nucleotide triplets with amino acids using an extensive dataset of protein–DNA complexes, and developed two support vector machine (SVM) models that predict protein-binding nucleotides from sequence data alone. One SVM model predicts protein-binding nucleotides using DNA sequence data alone, and the other SVM model predicts protein-binding nucleotides using both DNA and protein sequences. In a 10-fold cross-validation with 1519 DNA sequences, the SVM model that uses DNA sequence data only predicted protein-binding nucleotides with an accuracy of 67.0%, an *F*-measure of 67.1%, and a Matthews correlation coefficient (MCC) of 0.340. With an independent dataset of 181 DNAs that were not used in training, it achieved an accuracy of 66.2%, an *F*-measure 66.3% and a MCC of 0.324. Another SVM model that uses both DNA and protein sequences achieved an accuracy of 69.6%, an *F*-measure of 69.6%, and a MCC of 0.383 in a 10-fold cross-validation with 1519 DNA sequences and 859 protein sequences. With an independent dataset of 181 DNAs and 143 proteins, it showed an accuracy of 67.3%, an *F*-measure of 66.5% and a MCC of 0.329. Both in cross-validation and independent testing, the second SVM model that used both DNA and protein sequence data showed better performance than the first model that used DNA sequence data. To the best of our knowledge, this is the first attempt to predict protein-binding nucleotides in a given DNA sequence from the sequence data alone.

* *Corresponding author*. Tel.: +82 32 860 7388; fax: +82 32 863 4386.
E-mail address: khan@inha.ac.kr (K. Han).
http://dx.doi.org/10.1016/j.cmpb.2014.07.009

1. Introduction

The interactions between DNA and proteins comprise a fundamental role in many cellular processes [1]. For example, proteins that bind to specific regions of DNA act as transcription factors by activating or repressing gene expression of the DNA. Thus, identifying protein recognition parts in DNAs or DNA recognition parts in proteins will help understand a variety of cellular processes [2,3]. Protein–DNA interactions have been investigated in several theoretical or experimental studies, but the mechanism of protein–DNA interactions is not fully understood yet. Many studies on protein–DNA interactions have focused on investigating binding preferences of transcription factors in their DNA targets [4–7]. Predicting binding residues in protein sequences using machine learning methods have been carried out in several researches. BindN [3] uses a SVM to predict the RNA- or DNA-binding residues in protein sequences based on the chemical properties of amino acids. DNABindR [8] uses Naive Bayes classifier to predict the DNA-binding residues in proteins and DP-Bind [9] uses an SVM with a position specific scoring matrix (PSSM) and amino acid properties to predict DNA-binding residues in proteins. A recently published web server called PiDNA [10] computes the position weight matrix (PWM) for a given structure of protein–DNA complex, and predicts the protein–DNA interaction using the PWM suggested by the structure models with small energy changes. However, structure-based approaches such as PiDNA cannot be used to predict the interactions of protein and DNA with unknown structures.

While there are several studies for predicting DNA-binding residues in proteins, there have been few attempts to predict protein-binding nucleotides in DNA sequences using the sequence data alone. One of the reasons for this is that nucleotides show much less diverse interaction propensities than amino acids in protein–DNA interactions. Another reason is that there are only four types of nucleotides in DNA, whereas there are 20 types of amino acids in protein. For a sequence of length n, DNA has 4^n possible sequence patterns but protein has 20^n possible sequence patterns, which is 5^n fold larger. Thus, predicting the protein-binding nucleotides is more difficult than predicting the DNA-binding amino acids.

In this study, we analyzed recent structures of protein–DNA complexes and computed the interaction propensity between nucleotide triplets and amino acids using a support vector machine (SVM). We built two SVM models that use the interaction propensity to predict protein-binding sites in DNA sequences. One SVM model predicts protein-binding nucleotides from DNA sequence data alone, and the other SVM model predicts protein-binding nucleotides from DNA and protein sequence data. Both models take a single DNA sequence as input, so double-stranded DNA should be given to the model separately as single-stranded DNA. As shown later in this paper, experimental results with extensive datasets showed that the SVM model that uses both DNA and protein sequences exhibited better performance than the model that uses DNA sequences alone both in cross-validation and in independent testing. The rest of the paper presents the details of our approach and its experimental results.

2. Materials and methods

2.1. Definition of a binding site

The protein–DNA interaction data were extracted from the nucleic acid–protein interaction database (NPIDB) [11]. Among the three types of nucleic acid–protein interactions provided by NPIDB, we used hydrogen bonds (H bonds) and water bridges. Hydrophobic interaction data of NPIDB was not used for comparative analysis of two SVM models because one SVM model cannot match hydrophobic nucleic acid–protein interactions within a hydrophobic cluster [12].

For protein–DNA binding sites, the two types of DNA–protein interactions were incorporated into each sequence of protein–DNA complexes. For example, in a DNA chain of 10MH A4 (Adenine in the 4th position), G5, G7, C8 and A9 are involved in interactions with proteins. Nucleotides A4, G5, G7 and C8 bind to amino acids via H bonding interactions and water bridges, whereas A9 binds to amino acids via H bonding interaction only. In the example below, binding nucleotides are marked by the symbol '+' and non-binding nucleotides are marked by the symbol '-'.

	0	0	0	0	0	0	0	0	0	1	1	1
	1	2	3	4	5	6	7	8	9	0	1	2
DNA sequence:	G	T	C	A	G	C	G	C	A	T	G	G
H bonds:	-	-	-	+	+	-	+	+	+	-	-	-
Water bridges:	-	-	-	+	+	-	+	+	-	-	-	-
Binding sites:	-	-	-	+	+	-	+	+	+	-	-	-

2.2. Interaction propensity of nucleotide triplets

For the interaction propensity (IP) of a nucleotide triplet, we computed the propensity of the middle nucleotide in the triplet to interact with an amino acid. In the example below, the IP of the 5th nucleotide (C5) in the DNA sequence with the 10th amino acid (E10) in the protein sequence represents the binding preference of C in the triplet ACA for E.

	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7
DNA sequence:	C	T	C	A	C	A	C	G	T	G	G	G	A	C	T	A	G
Protein sequence:	M	K	R	E	S	H	K	H	A	E	Q	A	R	R	N	R	L
Binding pair:	(C-5, E-10)																

The interaction propensity IP_{ta} between the nucleotide triplet t and the amino acid a is defined by Eq. (1).

$$IP_{ta} = \sum_{\text{Binding Pair}} \frac{1}{\bar{HA}\cos(\angle DAH)} \cdot \frac{\sum_{i=\text{triplet}} N_i \cdot \sum_{j=AA} N_j}{N_t \cdot N_a \cdot \sum_{i=\text{triplet},j=AA} N_{ij}} \quad (1)$$

In Eq. (1), $\angle DAH$ is the donor–acceptor–hydrogen (D–A–H) angle. $\bar{HA}\cos(\angle DAH)$ is the projected length of the hydrogen–acceptor (H–A) on the donor–acceptor (D–A), $\sum N_{ij}$ is the total number of nucleotide triplets that bind to any amino acid, $\sum N_i$ is the total number of nucleotide triplets, $\sum N_j$ is the total number of amino acids, N_t is the number of nucleotide triplet t, and N_a is the number of amino acid a in the dataset. The reason for using the projected length of H–A is to consider the H–A distance and its relation with D–A together. There are $4^3 = 64$ nucleotide triplets and 20 amino acids, so we computed 1280 IPs between nucleotide triplets and amino acids.

2.3. Feature vector

Three types of features were used to predict protein-binding nucleotides: (1) global features of the DNA sequence, (2) local features of nucleotides, and (3) features of the interaction partner (i.e., protein features). The global features contain the entire DNA sequence information: the sequence length (L) and nucleotide composition (C). The nucleotide composition includes the number of adenine, cytosine, guanine, and thymine in the target DNA sequence. Thus, a feature vector always has one element for the DNA sequence length and four elements for the nucleotide composition as the global features.

The local features contain properties of a nucleotide in the DNA sequence: molecular mass (M), nucleotide pK_a (P), and nucleotide triplet IP with 20 amino acids. The first and the last nucleotides of a sliding window do not form a nucleotide triplet, so their IP values were set to 0 in the feature vector.

The partner features (A) contain information of the interacting protein sequence. This feature is computed by Eqs. (2) and (3) and represented as 20 elements for 20 types of amino acids in a feature vector.

$$P_{b\in\{20\text{ amino acids}\}} = \sum_{i,b_i=b}^{\text{sequence length}} \text{Normalized Position}\,(b_i) \tag{2}$$

$$\text{Normalized Position}\,(i) = \frac{\text{Position}\,(i)}{\text{Sequence Length}} \tag{3}$$

As mentioned earlier, we built two support vector machine (SVM) models to predict protein-binding nucleotides from sequence data only. For the SVM model that predicts protein-binding nucleotides using DNA sequence data only, the protein features (feature A) was not included in its feature vector. For the other SVM model that predicts protein-binding nucleotides using both DNA and protein sequences, the protein features (feature A) was encoded in its feature vector.

When we include the protein features in a feature vector using a window of 9 nucleotides, the feature vector contains 9 nucleotide triplets from $T(i-4)$ to $T(i+4)$ (see Fig. 1 for the structure of a feature vector). 5 global feature elements (L, C_A, C_C, C_G, C_T) and 20 elements for the protein partner ($A_R, \ldots, A_V$) are encoded once for a given pair of DNA and protein sequences. 22 local feature elements (M, P and 20 IPs) are encoded for 9 nucleotides (for $T(i-4)$ and $T(i+4)$, 0 are replaced for the triplet IPs). Thus, a feature vector representing a window of 9 overlapping nucleotide triplets has a total of 223 (=5 + 20 + 198) feature elements. If we do not consider the protein features, 20 elements are excluded. Thus, a feature vector contains 203 (=5 + 198) elements within it.

2.4. Performance measures

The performance of the prediction was evaluated with respect to seven measures: sensitivity (Sn, recall), specificity (Sp), accuracy (Ac), positive predictive value (PPV, precision), negative predictive value (NPV), F-measure (F), and Matthews correlation coefficient (MCC). The true positives (TP) are binding nucleotides that are correctly predicted as binding nucleotides, the true negatives (TN) are non-binding nucleotides that are predicted as non-binding nucleotides, the false positives (FP) are non-binding nucleotides that are incorrectly predicted as binding nucleotides, and the false negatives (FN) are binding nucleotides that are incorrectly predicted as non-binding nucleotides.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{6}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{7}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN}+\text{FN}} \tag{8}$$

$$F = \frac{2\times\text{Sn}\times\text{Sp}}{\text{Sn}+\text{Sp}} \tag{9}$$

$$\text{MCC} = \frac{(\text{TP}\times\text{TN})-(\text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{10}$$

Sensitivity is the ratio of correctly predicted binding nucleotides to actual binding nucleotides. Specificity is the ratio of correctly predicted non-binding nucleotides to actual non-binding nucleotides. Accuracy is the ratio of correctly predicted nucleotides to all nucleotides. Positive predictive value (PPV) measures the ratio of correctly predicted binding nucleotides to all nucleotides that are predicted as binding. Negative predictive value (NPV) measures the ratio of correctly predicted non-binding nucleotides to all nucleotides that are predicted as non-binding.

3. Results and discussion

3.1. Datasets

We obtained the structure data of protein–DNA complexes which were determined by X-ray crystallography with a resolution of 3.0 Åor better from the Protein Data Bank (PDB) [13]. As of May 2013, there were a total of 1691 protein–DNA complexes, and these complexes were used to compute the interaction propensity (IP) of nucleotide triplets with amino acids, which is defined by Eq. (1).

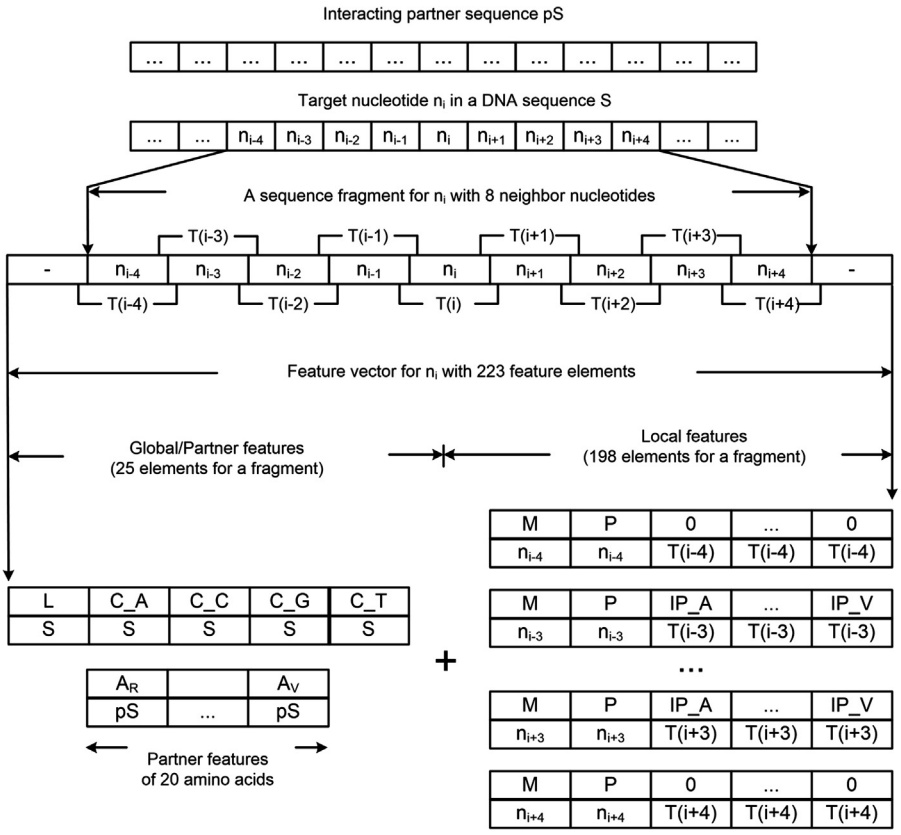


Fig. 1 – The structure of a feature vector with a window of 9 nucleotides. The structure of a feature vector with a window of 9 nucleotides, which covers 9 overlapping nucleotide triplets: *T*(i – 4), *T*(i – 3), . . ., *T*(i + 4). 5 global feature elements (L, C_A, C_C, C_G, C_T) and 20 protein feature elements (A_R, . . ., A_V) are encoded once for a pair of DNA and protein sequences. 22 local feature elements (M, P and 20 IPs) are encoded for each of the 9 nucleotides. For the IP of the ending nucleotide triplets, *T*(i – 4) and *T*(i + 4), of a window, 0 is encodes since their IP values are not defined. The number of elements in a feature vector is 223 (=5 + 20 + 198) for a window of 9 nucleotides. When the 20 protein features are not used, the number of a feature vector becomes 203 (=5 + 198).

The binding information of the 1691 protein–DNA complexes was obtained from NPIDB [11]. From the 1691 protein–DNA complexes, we extracted 5346 protein–DNA sequence pairs, which contain 1700 DNA sequences and 1002 protein sequences. By running CD-HIT [14] on the 1700 DNA sequences, we constructed a test dataset with 181 DNA sequences (hereafter called D250) whose sequence similarity is lower than 80% with others. The remaining 1519 DNA sequences were used to construct a training dataset (called D5096). The D250 dataset contained 181 DNA sequences and 143 protein sequences. The D5096 dataset contained 1519 DNA sequences and 859 protein sequences.

Our training dataset contains 202 transcription factors (23.5%) and other proteins. Furthermore, test dataset contains 44 transcription factors (30.8%) and other proteins. Thus, our models can predict not only transcription factor binding sites, but also other protein–DNA binding sites. The process of constructing the datasets is explained in Fig. 2.

We applied the feature-based redundancy removal method (F-method) [15] to the D5096 dataset to build two different training datasets. For the SVM model that uses DNA sequence data only, we constructed a training dataset TD1. Likewise, for the SVM model that uses both DNA and protein sequences, we constructed a training dataset TD2. Using TD1 and TD2, we tried several different window sizes from 11 to 39.

3.2. *Support vector machines*

SVM has been used to solve many prediction problems in bioinformatics. Shi et al. [16], for example, used SVM to predict protein phosphorylation sites. We used the library for SVM (LIBSVM) [17] to construct SVM models with the radial basis function (RBF) as a kernel. When training the SVM models, two parameter values must be adjusted: C and γ. The parameter C trades off misclassification of training data against simplicity of the decision surface. The parameter γ represents the width of the RBF. All the results shown in this paper were obtained with $C = 10$ and $\gamma = 1/\#$feature vectors for TD1 and TD2. As weights of positive and negative feature vectors, we used w_1 (weight of positive feature vectors) = 1.1 and w_{-1} (weight of negative feature vectors) = 1 for TD1 and w_1 (weight of positive feature vectors) = 1.45 and w_{-1} (weight of negative feature vectors) = 1 for TD2.

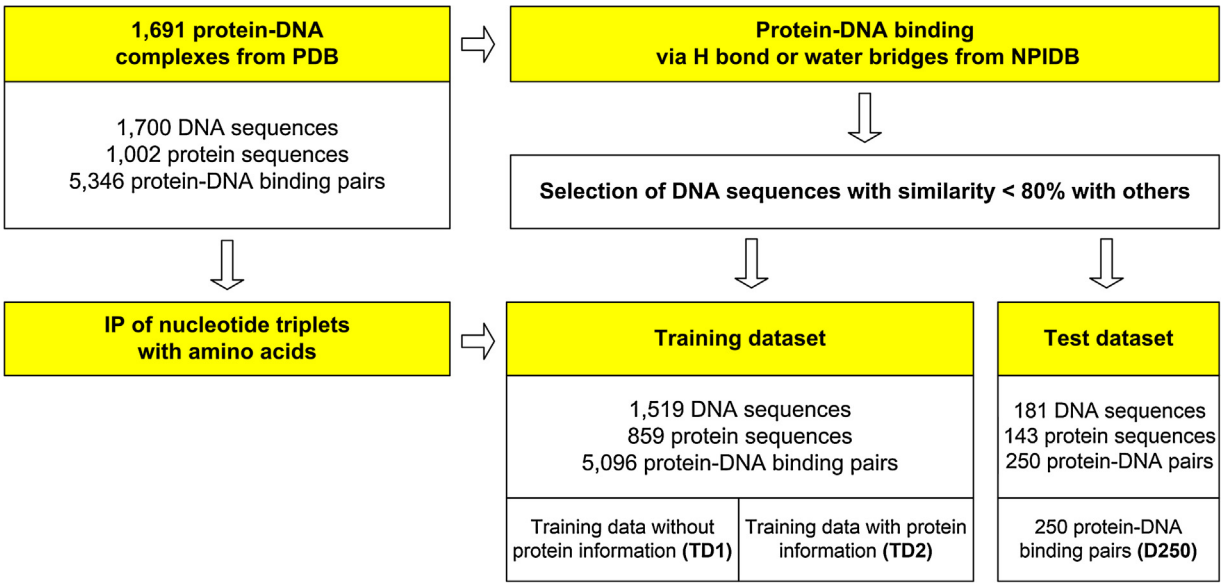


Fig. 2 – Construction of datasets for SVM models. The structure data of protein–DNA complexes that were determined by X-ray crystallography with a resolution of 3.0 Åor better were obtained from PDB. As of May 2013, there were a total of 1691 protein–DNA complexes, and these complexes were used to compute the interaction propensity (IP) of nucleotide triplets with amino acids. The 1691 protein–DNA complexes contained 5346 protein–DNA binding pairs, and the 5346 protein–DNA binding pairs were used to construct training and test datasets used in this study. The 5346 protein–DNA binding pairs contain 1700 DNA sequences and 1002 protein sequences. From the 1700 DNA sequences, a test dataset called D250 was first constructed by selecting 181 DNA sequences whose sequence similarity is lower than 80% with others. The remaining 1519 DNA sequences were used to construct a training dataset D5096. The D250 dataset contained 181 DNA sequences and 143 protein sequences. The D5096 dataset contained 1519 DNA sequences and 859 protein sequences.

3.3. *Performance of two prediction models*

For the SVM model that uses DNA sequence data only, we tried various window sizes, and Table 1 exhibited the prediction performance of the model with different window sizes from 11 to 39 nucleotides. In 10-fold cross validation, the window of 35 nucleotides exhibited the best performance (accuracy of 67.0% and MCC of 0.340). When the SVM model was tested on the 181 DNA sequences of D250, the best performance was observed with a window of 31 nucleotides (accuracy of 66.2% and MCC of 0.324).

On the other hand, Table 2 shows the result of the SVM model that uses both DNA and protein sequences. As shown in Table 2, a window of 35 nucleotides showed the best performance in the cross validation with TD2 (accuracy of 69.6% and MCC of 0.383). When the SVM model was tested on the 181 DNA sequences and 143 protein sequences of D250, it showed the best performance with a window of 31 nucleotides in most measures (accuracy of 67.3% and MCC of 0.329).

As the window size increases (testing on D250), MCC tends to increase but does not increase any more with a window of size 31 or larger. This is because a larger window includes more null nucleotides at both ends of the window. Details of testing of the SVM models with D250 are available in Additional file 1.

The disparity between the positive predictive value (PPV) and the negative predictive value (NPV) in Tables 1 and 2 can be explained as follows. The 10-fold cross validation with TD1 using different window sizes resulted in little difference (less than 10%) between PPV and NPV (Table 1). However, testing the model on D250 using a window of 31 nucleotides showed a difference of 12.4% between PPV and NPV. The 10-fold cross validation with TD2 using a window of 31 nucleotides showed a difference of 20.2% between PPV and NPV (Table 2), and testing the model on D250 using a window of 31 nucleotides showed a difference of 16.2% between PPV and NPV. The SVM model that uses both DNA and protein sequences predicts non-binding nucleotides more accurately than that uses DNA sequence only. Thus, a prediction model that uses both DNA and protein sequences resulted in a larger difference between PPV and NPV than that uses DNA sequence only.

Our prediction method is not restricted to DNA-binding sites of transcription factors, so both training and test datasets used in our study include transcription factors and other proteins as well as binding partners of DNA. In fact, 202 (23.5%) out of the 859 protein sequences in the training dataset D5096 are transcription factors, and 44 (30.8%) out of the 143 protein sequences in the test dataset D250 are transcription factors. Experimental results of the prediction models showed that they can predict not only DNA-binding sites of transcription factors but DNA-binding sites of other types of proteins.

To test our prediction model on DNA sequences with unknown structures, we tested it on three datasets: (1) data from JASPAR [18], (2) random DNA sequences, and (3) the *Eschrichia coli* genome with transcription factor binding sites removed. First, we extracted 163 transcription factors and their DNA-binding sites from the JASPAR database [18]. PDB contains no protein–DNA complex that includes any of the 163 transcription factors, and the 163 transcription factors

Table 1 – The performance of a SVM model that uses DNA sequence data only. A total of 5 features (2 global features and 3 local features) were used.

WS	Ac	Sn	Sp	F	PPV	NPV	AUC	MCC
10-Fold cross validation with TD1								
11	62.9%	65.1%	60.9%	62.9%	58.9%	67.0%	0.6803	0.2589
15	63.9%	65.4%	62.7%	64.0%	60.0%	67.9%	0.6965	0.2798
19	65.4%	64.8%	66.0%	65.4%	61.7%	68.9%	0.7082	0.3069
23	65.9%	66.4%	65.5%	66.0%	61.7%	70.0%	0.7188	0.3179
27	66.4%	67.3%	65.7%	66.5%	61.2%	70.9%	0.7256	0.3283
31	66.7%	67.7%	65.9%	66.8%	61.8%	71.4%	0.7292	0.3341
35	67.0%	67.9%	66.4%	67.1%	62.1%	71.8%	0.7304	0.3401
39	67.0%	67.9%	66.2%	67.1%	61.9%	71.9%	0.7305	0.3396
Testing on D250								
11	63.8%	65.0%	62.8%	63.9%	57.3%	70.0%	0.6859	0.2761
15	64.4%	65.2%	63.7%	64.4%	58.0%	70.5%	0.6931	0.2869
19	64.7%	64.3%	64.9%	64.6%	58.4%	70.3%	0.7049	0.2902
23	65.3%	67.2%	63.9%	65.5%	58.8%	71.7%	0.7090	0.3078
27	65.5%	66.9%	64.5%	65.7%	59.1%	71.7%	0.7162	0.3108
31	66.2%	67.3%	65.3%	66.3%	59.9%	72.3%	0.7174	0.3239
35	65.1%	66.3%	64.2%	65.2%	58.7%	71.3%	0.7126	0.3025
39	65.2%	66.4%	64.3%	65.3%	58.8%	71.4%	0.7108	0.3044

WS: window size, Ac: accuracy, Sn: sensitivity, Sp: specificity, F: F-measure, PPV: positive predictive value, NPV: negative predictive value, AUC: the area under the ROC curve, MCC: Matthews correlation coefficient.

and their binding data were not used in training the prediction model at all. For the 163 transcription factors, the model predicted 98.5% of their DNA-binding sites correctly. The 163 JASPAR IDs and prediction results are available in Additional file 2.

For the second test dataset, we generated 1000 random DNA sequences of 40–100 nucleotides with the same nucleotide composition using FaBox [19]. We tested our model and evaluated its specificity with the assumption that the random DNA sequences contain no protein-binding sites. The specificity of the model with the random sequences was 63.8%, which is slightly lower than those with actual data (Table 1). Additional file 3 shows the 1000 random DNA sequences and prediction results.

For the third test dataset, we removed all known transcription factor binding sites [20] from the *E. coli* genome. The remaining *E. coli* genome was assumed to contain no protein-binding sites since is no evidence of protein binding at present.

Table 2 – The performance of a SVM model that uses both DNA and protein sequences. A total of 6 features (2 global features, 3 local features, and the protein features) were used.

WS	Ac	Sn	Sp	F	PPV	NPV	AUC	MCC
10-Fold cross validation with TD2								
11	67.6%	69.4%	66.4%	67.9%	55.9%	77.9%	0.7453	0.3485
15	68.3%	67.8%	68.6%	68.2%	57.1%	77.6%	0.7497	0.3556
19	69.1%	66.9%	70.4%	68.6%	58.1%	77.6%	0.7521	0.3648
23	69.4%	68.1%	70.2%	69.2%	58.4%	78.2%	0.7576	0.3748
27	69.7%	69.1%	70.2%	69.6%	58.7%	78.7%	0.7601	0.3829
31	69.7%	69.3%	70.0%	69.6%	58.6%	78.8%	0.7605	0.3826
35	69.6%	69.7%	70.0%	69.6%	58.4%	79.0%	0.7603	0.3832
39	69.4%	69.9%	69.1%	69.5%	58.1%	79.0%	0.7591	0.3802
Testing on D250								
11	64.0%	64.1%	64.0%	64.0%	54.0%	73.0%	0.6901	0.2754
15	64.6%	61.7%	66.5%	64.0%	54.8%	72.4%	0.6950	0.2768
19	66.1%	62.9%	68.1%	65.4%	56.6%	73.6%	0.7043	0.3061
23	65.7%	63.3%	67.4%	65.3%	56.2%	73.5%	0.7106	0.3017
27	67.1%	63.9%	69.3%	66.5%	57.9%	74.4%	0.7120	0.3275
31	67.3%	63.5%	69.8%	66.5%	58.1%	74.3%	0.7160	0.3288
35	67.1%	64.1%	69.0%	66.4%	57.8%	74.4%	0.7144	0.3264
39	67.0%	63.4%	69.3%	66.2%	57.7%	74.1%	0.7107	0.3229

WS: window size, Ac: accuracy, Sn: sensitivity, Sp: specificity, F: F-measure, PPV: positive predictive value, NPV: negative predictive value, AUC: the area under the ROC curve, MCC: Matthews correlation coefficient.

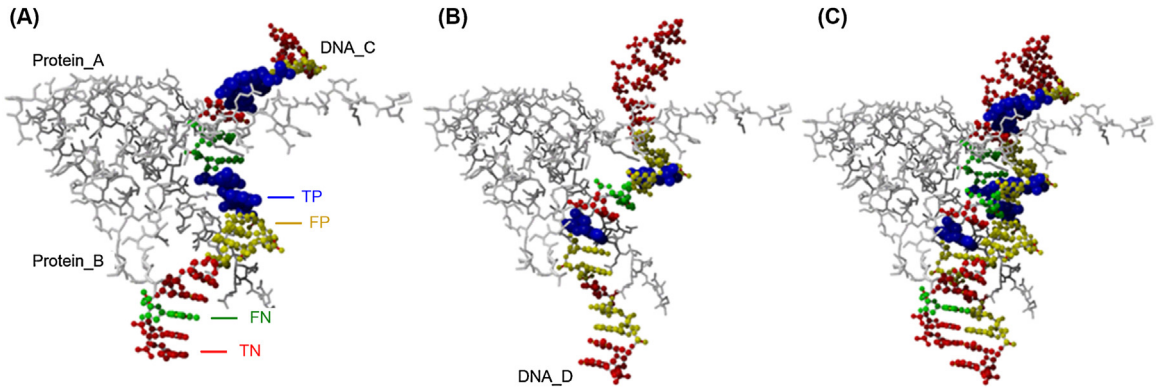


Fig. 3 – Example of predicting protein-binding sites in DNA chains using DNA sequence data only. Gray sticks represent protein sequences of a protein–DNA complex 1AN4. (A) In DNA chain C, the model predicted 4 binding nucleotides (TP) and 8 non-binding nucleotides (TN) correctly. 6 nucleotides are actually non-binding but were predicted as binding (FP). 3 nucleotides are actually binding but were predicted as non-binding (FN). (B) In DNA chain D, the model predicted 2 binding nucleotides (TP) and 10 non-binding nucleotides (TN) correctly. 8 nucleotides are actually non-binding but were predicted as binding (FP). 1 nucleotides are actually binding but were predicted as non-binding (FN). (C) Superimposed structure showing protein-binding sites in DNA chains C and D. TP: true positives (blue), TN: true negatives (red), FP: false positives (yellow), FN: false negatives (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

From the remaining *E. coli* genome, we randomly extracted 1000 DNA sequences of 40–100 nucleotides and tested our model on them. The model showed the specificity of 66.3%, which is close to the best specificity of 65.3% (testing on D250 with WS = 31 in Table 1). Additional file 4 shows the 1000 *E. coli* sequences and prediction results.

3.4. *Example of protein-binding nucleotides predicted by two models*

Fig. 3 shows an example of predicting protein-binding sites in DNA chains using DNA sequence data only. Fig. 4 shows the protein-binding sites in DNA chains predicted by another SVM model which uses both DNA and protein sequences. Apparently, the protein-binding sites predicted by two models are different since one of them uses additional information on the binding partner of DNA for predicting binding sites. While the model that uses DNA sequence only predicted 6 (=4 + 2) binding nucleotides and 18 (=8 + 10) non-binding nucleotides correctly, the other model that uses both DNA and protein sequences predicted 5 (=3 + 2) binding nucleotides and 20 (=9 + 11) non-binding nucleotides correctly. The model that uses both DNA and protein sequences has 2 (=14 − 12) fewer false positives (FP) and 1 (=5 − 4) more false

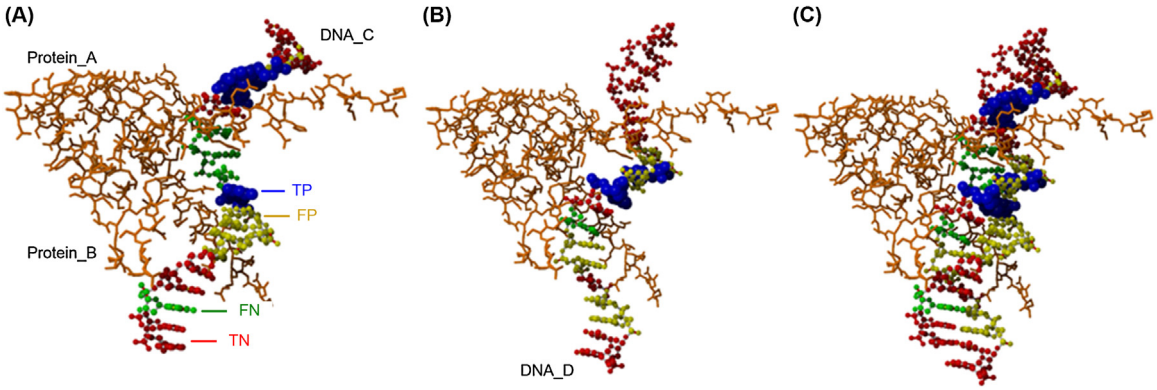


Fig. 4 – Example of predicting of protein-binding sites in DNA chains using both DNA and protein sequences. The two protein chains A and B (represented by orange sticks) of a protein–DNA complex 1AN4 have the same amino acid sequence. (A) In DNA chain C, the model predicted 3 binding nucleotides (TP) and 9 non-binding nucleotides (TN) correctly. 5 nucleotides are actually non-binding but were predicted as binding (FP). 4 nucleotides are actually binding but were predicted as non-binding (FN). (B) In DNA chain D, the model predicted 2 binding nucleotides (TP) and 11 non-binding nucleotides (TN) correctly. 7 nucleotides are actually non-binding but were predicted as binding (FP). 1 nucleotides are actually binding but were predicted as non-binding (FN). (C) Superimposed structure showing protein-binding sites in DNA chains C and D. TP: true positives (blue), TN: true negatives (red), FP: false positives (yellow), FN: false negatives (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

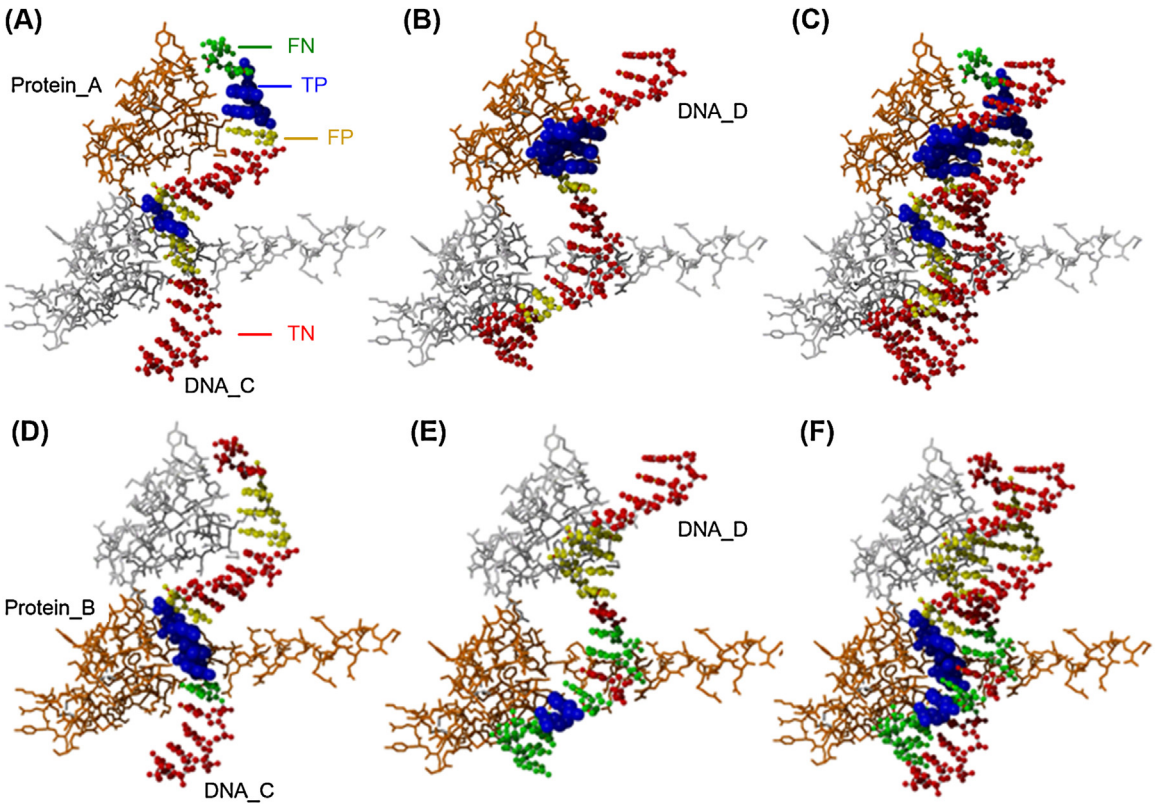


Fig. 5 – Example of predicting of binding sites in DNA chains C and D with different proteins A and B of 2NLL. Orange sticks represent the protein of 2NLL that is specified as a binding partner of DNA, and gray sticks represent the protein that is not selected as a binding partner. (A) When a protein chain A was given as a binding partner of a DNA chain C, the model predicted 3 binding nucleotides (TP) and 9 non-binding nucleotides (TN) correctly. 4 nucleotides are non-binding but were predicted as binding (FP). 2 nucleotides are actually binding but were predicted as non-binding (FN). (B) When a protein chain A was given as a binding partner of a DNA chain D, it predicted 3 binding nucleotides and 13 non-binding nucleotides correctly. 2 nucleotide is binding but was predicted as non-binding. (C) Superimposed structure that shows binding sites of DNA chains C and D with protein chain A. (D) When a protein chain B was given as a binding partner of a DNA chain C, it predicted 3 binding nucleotides and 10 non-binding nucleotides correctly. 4 nucleotides are actually non-binding but were predicted as binding. 1 nucleotide is actually binding but was predicted as non-binding. (E) When a protein chain B was given as a binding partner of a DNA chain D, it predicted 1 binding nucleotide and 6 non-binding nucleotides correctly. 4 nucleotides are non-binding but were predicted as binding. 7 nucleotides are binding but were predicted as non-binding. (F) Superimposed structure that shows the binding sites in DNA chains C and D with protein chain B. TP: true positives (blue), TN: true negatives (red), FP: false positives (yellow), FN: false negatives (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

negatives (FN). Thus, the model that uses both DNA and protein sequences had one more number of TP+TN, and contained one fewer number in FP+FN. The model that uses both DNA and protein sequences was slightly better than the model that uses DNA sequences only in predicting non-binding nucleotides.

Fig. 5 shows another example of predicting of binding sites in DNA chains with different protein partners. Even for a same DNA sequence, protein-binding sites in the DNA sequence can be changed when its binding partner changes. The average prediction results of DNA chain C with protein A and DNA chain D with protein A of 2NLL are PPV 51.5% and NPV 90.9%, as well as the average prediction results of DNA chain C with protein B and DNA chain D with protein B are PPV 31.5% and NPV 68.6%. On the other hand, the average prediction results of DNA chain C and DNA chain D only are PPV 66.7% and NPV 66.7%. Likewise, the model that uses both DNA and protein sequences was better than model that uses DNA sequences only in predicting non-binding nucleotides. Also, these prediction results are available on Additional file 1.

4. Conclusions

We computed the interaction propensity (IP) of nucleotide triplets with amino acids using an extensive dataset

of protein–DNA complexes, and developed two support vector machine (SVM) models that predict protein-binding nucleotides from sequence data alone. One SVM model predicts protein-binding nucleotides using DNA sequence data alone, and the other SVM model predicts protein-binding nucleotides using both DNA and protein sequences. In a 10-fold cross-validation with 1519 DNA sequences, the SVM model that uses DNA sequence data only predicted protein-binding nucleotides with an accuracy of 67.0%, an *F*-measure of 67.1%, and a Matthews correlation coefficient (MCC) of 0.340. With an independent dataset of 181 DNAs that were not used in training, it achieved an accuracy of 66.2%, an *F*-measure 66.3% and a MCC of 0.324. Another SVM model that uses both DNA and protein sequences achieved an accuracy of 69.6%, an *F*-measure of 69.6%, and a MCC of 0.383 in a 10-fold cross-validation with 1519 DNA sequences and 859 protein sequences. With an independent dataset of 181 DNAs and 143 proteins, it showed an accuracy of 67.3%, an *F*-measure of 66.5% and a MCC of 0.329.

A prediction model that uses both DNA and protein sequences exhibited improved and reliable performance than using DNA sequences alone when predicting the protein-binding sites in a DNA sequence. Unlike structure-based approaches, our method does not assume the structure of DNA is known. If the sequence data of interacting protein is available, protein-binding nucleotides can be predicted better using the protein sequence data as well. Our method is not restricted to DNA-binding sites of transcription factors. Both transcription factors and other proteins were included as binding partners of DNA in training and test datasets of our prediction models. It will be useful to find protein-binding sites in DNA when its structure is not known and the sequence is the only information available. To the best of our knowledge, this is the first attempt to predict protein-binding DNA nucleotides with sequence data alone.

Conflicts of interest

None declared.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2012R1A1A3011982) and in part by the Ministry of Education (2010-0020163) and Inha University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cmpb.2014.07.009.

REFERENCES

[1] X.M. Ding, X.Y. Pan, C. Xu, H.B. Shen, Computational prediction of DNA–protein interactions: a review, Curr. Comput. Aided Drug Des. 6 (3) (2010) 197–206.
[2] S.Y. Ho, F.C. Yu, C.Y. Chang, H.L. Huang, Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method, Biosystems 90 (1) (2007) 234–241.
[3] L.J. Wang, S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, Nucleic Acids Res. 34 (W1) (2006) W243–W248.
[4] Z. Qian, L. Lu, X. Liu, Y.-D. Cai, Y. Li, An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization, Bioinformatics 23 (18) (2007) 2449–2454.
[5] S. Yang, H.K. Yalamanchili, X. Li, K.-M. Yao, P.C. Sham, M.Q. Zhang, J. Wang, Correlated evolution of transcription factors and their binding sites, Bioinformatics 27 (21) (2011) 2972–2978.
[6] G. Zheng, Q. Liu, G. Ding, C. Wei, Y. Li, Towards biological characters of interactions between transcription factors and their DNA targets in mammals, BMC Genomics 13 (1) (2012) 388.
[7] P. Athanasiadis, A. Malousi, S. Kouidou, N. Maglaveras, Gremet: an integrative tool for the prediction of mutation effects on gene regulation, Comput. Methods Programs Biomed. 111 (1) (2013) 214–219.
[8] C.H. Yan, M. Terribilini, F. Wu, R.L. Jernigan, D. Dobbs, V. Honavar, Predicting DNA-binding sites of proteins from amino acid sequence, BMC Bioinform. 7 (2006), http://dx.doi.org/10.1186/1471-2105-7-262.
[9] S. Hwang, Z.K. Gou, I.B. Kuznetsov, DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins, Bioinformatics 23 (5) (2007) 634–636.
[10] C.K. Lin, C.Y. Chen, PiDNA: predicting protein–DNA interactions with structural models, Nucleic Acids Res. 41 (W1) (2013) W523–W530.
[11] D.D. Kirsanov, O.N. Zanegina, E.A. Aksianov, S.A. Spirin, A.S. Karyagina, A.V. Alexeevski, NPIDB: nucleic acid–protein interaction database, Nucleic Acids Res. 41 (D1) (2013) D517–D523.
[12] A. Alexeevski, S. Spirin, D. Alexeevski, O. Klychnikov, A. Ershova, M. Titov, A. Karyagina, CluD, a program for determination of hydrophobic clusters in 3D structures of protein and protein–nucleic acids complexes, Biophysics 48 (S1) (2004) 146–150.
[13] P.W. Rose, B. Beran, C.X. Bi, W.F. Bluhm, D. Dimitropoulos, D.S. Goodsell, A. Prlic, M. Quesada, G.B. Quinn, J.D. Westbrook, J. Young, B. Yukich, C. Zardecki, H.M. Berman, P.E. Bourne, The RCSB protein data bank: redesigned web site and web services, Nucleic Acids Res. 39 (D1) (2011) D392–D401.
[14] Y. Huang, B.F. Niu, Y. Gao, L.M. Fu, W.Z. Li, CD-HIT suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (5) (2010) 680–682.
[15] S. Choi, K. Han, Prediction of RNA-binding amino acids from protein and RNA sequences, BMC Bioinform. 12 (S13) (2011), http://dx.doi.org/10.1186/1471-2105-12-S13-S7.
[16] Y. Shi, B. Yuan, G. Lin, D. Schuurmans, Protein phosphorylation site prediction via feature discovery support vector machine, Tsinghua Sci. Technol. 17 (6) (2012) 638–644.
[17] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[18] A. Mathelier, X. Zhao, A.W. Zhang, F. Parcy, R. Worsley-Hunt, D.J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin,

W.W. Wasserman, Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles, Nucleic Acids Res. 42 (D1) (2014) D142–D147.

[19] P. Villesen, Fabox: an online toolbox for fasta sequences, Mol. Ecol. Notes 7 (6) (2007) 965–968.

[20] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muiz-Rascado, J.S. Garca-Sotelo, V. Weiss, H. Solano-Lira, I. Martnez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernndez, K. Alquicira-Hernndez, A. Lpez-Fuentes, L. Porrn-Sotelo, A.M. Huerta, C. Bonavides-Martnez, Y.I. Balderas-Martnez, L. Pannier, M. Olvera, A. Labastida, V. Jimnez-Jacinto, L. Vega-Alvarado, V. del Moral-Chvez, A. Hernndez-Alvarez, E. Morett, J. Collado-Vides, Regulondb v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, Nucleic Acids Res. 41 (D1) (2013) D203–D213.